THE SELECTIVE USE OF GAZE IN AUTOMATIC SPEECH RECOGNITION

by

AO SHEN

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Electronic, Electrical and Computer Engineering
College of Engineering and Physical Sciences
The University of Birmingham
August 2013

Abstract

The performance of automatic speech recognition (ASR) degrades significantly in natural environments compared to in laboratory assessments. Being a major source of interference, acoustic noise affects speech intelligibility during the ASR process. There are two main problems caused by the acoustic noise. The first is the speech signal contamination. The second is the speakers' vocal and non-vocal behavioural changes. These phenomena elicit mismatch between the ASR training and recognition conditions, which leads to considerable performance degradation. To improve noise-robustness, exploiting prior knowledge of the acoustic noise in speech enhancement, feature extraction and recognition models are popular approaches. An alternative approach presented in this thesis is to introduce eye gaze as an extra modality. Eye gaze behaviours have roles in interaction and contain information about cognition and visual attention; not all behaviours are relevant to speech. Therefore, gaze behaviours are used selectively to improve ASR performance. This is achieved by inference procedures using noise-dependant models of gaze behaviours and their temporal and semantic relationship with speech. 'Selective gaze-contingent ASR' systems are proposed and evaluated on a corpus of eye movement and related speech in different clean, noisy environments. The best performing systems utilise both acoustic and language model adaptation and show a statistically significant improvement in word error rates. The work highlights a methodology for using gaze and loosely coupled non-verbal modalities selectively to achieve noise-robust ASR.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to Dr. Neil Cooke, who not only provided the professional supervision to this work, but also encouraged me through all these years. I want to thank Mrs Mary Winkles for her counsel; Colleagues for their help, comments and advices.

Finally I wish to thank my parents Yide Shen and Yongchun Qiu, for their steadfast support and love.

CONTENTS

1 Introduction **4**

1.1 Multimodal Interaction 4

1.2 Automatic Speech Recognition 6

1.3 Eye Tracking 7

1.4 Gaze-contingent ASR 8

1.5 'Selective Use' of Events in Multimodal Integration 9

1.6 'Selective' Gaze-contingent ASR 10

1.7 Research Questions and Methodology 11

1.8 Thesis Structure 12

2 Gaze, Speech, and Multimodal Systems **14**

2.1 Multimodal Interaction 14

2.1.1 Multimodal systems 14

2.1.2 Use of modalities 16

2.1.3 Robustness of recognition 17

2.1.4 Context awareness 18

2.2 Gaze in Multimodal Systems 19

2.2.1 Selective use of gaze 19

2.2.2 Gaze features 21

2.2.3 Visual attention 22

2.2.4 The meaning or role of gaze 23

2.2.5 Machine learning approaches in gaze analysis 25

2.2.6 Gaze and speech in multimodal systems . . . 26
2.3 Acoustic Noise and Its Impact on ASR Performance . . . 28
2.3.1 Automatic speech recognition . . . 28
2.3.2 Effect of noise in ASR performance . . . 30
2.3.3 Acoustic Lombard effect . . . 31
2.3.4 Strategies for noise-robust ASR . . . 32
2.3.5 Lombard effect and ASR . . . 35
2.4 Multimodal ASR . . . 37
2.4.1 Types of multimodal ASR . . . 37
2.4.2 Handling noise in multimodal ASR . . . 38
2.4.3 Gaze-contingent ASR . . . 39
2.5 Summary . . . 41

3 Integration of Gaze and Speech in ASR 42
3.1 Architecture . . . 42
3.2 Taxonomy of Gaze Roles for ANI and VAI . . . 43
3.3 Coupling between Gaze And Speech Events . . . 46
3.4 Mutual Information And Its Utilisation in Acoustic Noise Inference (ANI) 49
3.5 Visual Attention Inference (VAI) . . . 52
3.6 Generalising Speech to Other Modalities . . . 53
3.7 Summary . . . 53

4 A Corpus of Gaze And Speech in Acoustic Noise 54
4.1 Motivation for Collecting the ES-N Corpus . . . 55
4.2 Wizard-of-Oz Simulation . . . 56
4.2.1 WoZ as a research method . . . 56
4.2.2 Motivation . . . 57
4.2.3 Abstract system descriptions . . . 59
4.2.4 Data collection and pilot study . . . 59

4.3 Method . . . 60
4.3.1 Task . . . 60
4.3.2 WoZ implementation and system descriptions . . . 61
4.3.3 Experimental procedure for corpus collection . . . 63
4.3.4 Participants . . . 63
4.3.5 Apparatus . . . 64
4.3.6 Experiment design application . . . 65
4.4 Post Processing . . . 66
4.4.1 Synchronisation . . . 67
4.4.2 Speech transcription . . . 69
4.4.3 Calibration errors and quality assessment . . . 69
4.5 Main Data Collection . . . 72
4.5.1 Adding acoustic noise . . . 72
4.5.2 Session structure . . . 74
4.6 Summary . . . 74

5 Acoustic Noise Inference and The Gaze Lombard Effect 76

5.1 Motivation . . . 77
5.2 Assumption of Normality . . . 77
5.3 Pilot Study . . . 78
5.3.1 Tests for noise type comparison . . . 78
5.3.2 Pilot data results . . . 79
5.4 Acoustic Lombard Effect Analysis . . . 84
5.4.1 Speech data test performed . . . 84
5.4.2 Results . . . 85
5.5 Gaze Lombard Effect Analysis . . . 86
5.5.1 Gaze data test performed . . . 86
5.5.2 Results . . . 87
5.6 Acoustic Noise Inference . . . 97

5.6.1 Density estimation in MI calculation . . . 98
5.6.2 Test conducted . . . 101
5.6.3 MI results (Test 1 & 2) . . . 102
5.6.4 Discussions as to SVM experimental setup . . . 103
5.6.5 Noise classification results (Test 3 & 4) . . . 107
5.7 Summary . . . 110

6 Visual Attention Inference **112**
6.1 VAI Implementation for the ES-N Task . . . 112
6.2 Method . . . 116
6.3 Feature Selection for VAI . . . 118
6.3.1 Data labelling . . . 118
6.3.2 Feature extraction . . . 118
6.3.3 Feature normalisation . . . 119
6.3.4 Feature discriminability in no-noise condition . . . 120
6.3.5 Feature discriminability and dependency upon acoustic noise . . . 122
6.4 VAI Performance Test . . . 130
6.5 Results . . . 132
6.6 Summary . . . 134

7 Selective Gaze-Contingent ASR System **137**
7.1 ASR and Adaptation Overview . . . 137
7.1.1 ASR basics . . . 137
7.1.2 Language Model . . . 138
7.1.3 Selective gaze-contingent ASR architecture . . . 140
7.1.4 Gaze-based LM adaptation . . . 140
7.1.5 Cache-based LM adaptation . . . 142
7.1.6 Acoustic model adaptation . . . 143
7.2 Framework for Selective Gaze Integration . . . 144

7.2.1 Baseline LM construction using class-based model . . . 144
7.2.2 Cache-based LM adaptation . . . 145
7.2.3 Relevance function . . . 147
7.2.4 VAI implementation for relevance functions . . . 148
7.3 Evaluation Methodology . . . 149
7.3.1 Method . . . 149
7.3.2 Baseline LM and class construction . . . 149
7.3.3 Training the baseline ASR . . . 150
7.3.4 Parameter selection . . . 150
7.3.5 LM performance measure . . . 153
7.3.6 ASR performance measure . . . 153
7.3.7 N-best rescoring . . . 154
7.4 Tests Conducted . . . 155
7.4.1 Test 1: language model (LM) adaptation performance . . . 155
7.4.2 Test 2: acoustic model (AM) adaptation performance . . . 155
7.4.3 Test 3: VAI-based LM adaptation performance in ASR . . . 156
7.4.4 Test 4: selective gaze-Contingent ASR performance . . . 157
7.5 ASR Results . . . 157
7.5.1 LM adaptation perplexity performance (Test 1) . . . 157
7.5.2 AM adaptation WER performance (Test 2) . . . 159
7.5.3 ASR WER performance for evaluating the LM adaptation (Test 3) 161
7.5.4 Selective gaze-contingent ASR WER performance (Test 4) . . . 163
7.6 Summary . . . 164

8 Conclusion **167**

8.1 Contributions . . . 168
8.1.1 A formalised framework for measuring the coupling between modalities . . . 168
8.1.2 A working taxonomy of gaze roles . . . 169

8.1.3 The ES-N corpus recorded in different acoustic noise 169
8.1.4 The 'gaze Lombard effect' and the dependency of the gaze-speech relationship on acoustic noise 170
8.1.5 A cache-based LM adaptation approach using class-based model . . 170
8.1.6 A noise-robust, selective gaze-contingent ASR 171
8.2 Recommendations for Future Research 172
8.2.1 Corpus and General Framework 172
8.2.2 Noise Condition and Affective State 173
8.2.3 Acoustic Noise Inference Using Statistical Gaze Information 174
8.2.4 Lombard Effect in Different Modalities and Variability between People 175
8.2.5 ASR Vocabulary, Segmentation and Language Model 175
8.3 Summary 176

Appendix A: publication **177**

List of References **178**

LIST OF TABLES

4.1 SR Research Eyelink 2 eye-tracker technical specifications 65
4.2 A summary of the sessions collected. 75

5.1 Fixation Duration and Saccade Length across three noise types. 81
5.2 Speech characteristics in 4 noise conditions. 85
5.3 Breakdown in the average speech rate (wpm) for the 7 participants. 86
5.4 The number of the fixations with the means and the standard deviations across 4 noise conditions. 89
5.5 The frequency statistics of the fixation events 'during speech' and 'during silence'. 94
5.6 The mean value of the fixation duration 'during speech' and 'during silence' across 4 noise conditions. 94
5.7 Breakdown in the relative mean fixation duration change from N0 to N3 for the 7 participants. 97
5.8 SVM classifier performance for the four noise conditions (n=377). 109
5.9 Confusion matrix of instance numbers for noise classification results given in table 5.8. 109
5.10 Two-class SVM classifier performance for no noise and 15dB noise (n=184). 109
5.11 SVM classifier performance for 7 participants over all tasks (n=377). . . . 110

6.1 Features computed from fixation, saccade and pupil size 119
6.2 VAI input data frequency. 132
6.3 Classifier performance for gaze role inference in no noise environment. . . . 133
6.4 The performances before and after incorporating the coupling function f. . 135
6.5 The confusion matrix for no-noise condition before and after incorporating the coupling function. 135
6.6 The confusion matrix for noisy condition before and after incorporating the coupling function. 135

7.1 Language model perplexity results. 159
7.2 The adaptation results measured in WER. 160
7.3 The WER performances before and after the LM adaptations. 163
7.4 The breakdown of the baseline performance in terms of noise conditions and participants. 163
7.5 ASR system performances in terms of WER on gaze and speech data recorded in various acoustic noise. 164

LIST OF FIGURES

1.1 The hierarchical thesis structure. 12

3.1 The system design architecture for inference of acoustic noise condition and relevant visual attentions. 44
3.2 Proposed working gaze role taxonomy for this thesis. 47
3.3 Proposed coupling framework between information events in different modalities. 48

4.1 The spatial puzzle task used to elicit different gaze behaviours in acoustic noise. 61
4.2 A sample dialogue for the spatial puzzle task. 62
4.3 The set up of the eye-tracker system. 66
4.4 Hierarchical organisation of events used in the experiment design flow. . . . 67
4.5 A typical example of the Trail (Session) Level design flow. 68
4.6 An example of the phoneme-level time-aligned transcription. 69
4.7 Examples of the good eye-camera positions and unsatisfactory positions. . 71
4.8 Spatial Overlay View interface. 73
4.9 Temporal Overlay View interface. 73

5.1 Participants' fixation duration and saccade length in different noise types. . 80
5.2 The normality QQ-plot for fixation duration and saccade length in different noise types. 82
5.3 The MI results between speech and gaze across three noise types. 83
5.4 An example of captured gaze and speech features without and with environmental acoustic noise. 84
5.5 The histogram of fixations under 4 noise conditions and the normal distribution lines with a decreasing kurtosis across the noise condition. 88
5.6 The normality Q-Q plot of fixations under N0 and N1. 90
5.7 The normality Q-Q plot of fixations under N2 and N3. 91
5.8 The mean value of the fixation durations (ms) across 4 noise groups. 92
5.9 The mean value of the saccade length across 4 noise groups. 93
5.10 The distinction between the fixations 'during silence' and 'during speech' across 4 noise conditions. 95
5.11 Example of calculating two different measures of mutual information between gaze sequence G and speech sequence W for noise-inference. 99
5.12 The MI values based on the coupling of mediating attention (MA) and object naming (ON) respectively. 104
5.13 The classification process framework. 106

5.14 The definition of common classification evaluation metrics. 107
5.15 An example contour plot of classification accuracy. 108

6.1 TIVA, TOVA, and RVA events in the form of sequences for feature extraction. 114
6.2 The coupling function between a gaze event g_t and a speech event w_v. . . . 115
6.3 The overall fixation duration distribution before and after the nature log transform. 121
6.4 Difference in fixation duration distribution for each role before and after normalisation. 121
6.5 Summary statistics for normalised gaze features on a per-role basis without environmental acoustic noise. 122
6.6 The bar graph of the event fixation duration z-score. 124
6.7 The bar graph of the prior fixation duration z-score. 125
6.8 The bar graph of the post fixation duration z-score. 127
6.9 The bar graph of the prior saccade length z-score. 128
6.10 The bar graph of the post saccade length z-score. 129
6.11 The bar graph of the average pupil size change. 131
6.12 The weighted average AUC of the classifier for no-noise and noisy conditions. 134

7.1 A standard basic model of automatic speech recognition. 139
7.2 The implementation architecture of the selective gaze-contingent ASR. . . 141
7.3 Cache-based LM adaptation framework. 146
7.4 Surface plot showing WER as a function of WIP and LMSF. 152
7.5 The perplexity values of the three approaches. 158
7.6 The effect of LM adaptation in no-noise condition (N0). 162
7.7 The effect of LM adaptation in the noisiest condition (N3). 162

Abbreviations

AM Acoustic Model

ANN Artificial Neural Network

ANI Acoustic Noise Inference

ASR Automatic Speech Recognition

AUC Area Under Curve

BEEP British English Example Pronunciation

CA Context Awareness

CCA Cognitive Context Awareness

DAG Direct Acyclic Graph

ECA Environmental Context Awareness

EFD Event Fixation Duration

ES-N Eye-Speech-in-Noise Corpus

GUI Graphical User Interface

HCI Human-Computer Interaction

HMM Hidden Markov Model

HTK The Hidden Markov Model ToolKit

ICA Interactional Context Awareness

LM Language Model

LMSF Language Model Scale Factor

MA Mediating Attention

MAP Maximum A-Posteriori

MI Mutual Information

ML Machine Learning

MLLR Maximum Likelihood Linear Regression

ON Object Naming

PDF Probability Density Function

PRFD Prior Fixation Duration

PRSL Prior Saccade Length

PTFD Post Fixation Duration

PTSL Post Saccade Length

RVA Reactive Visual Attention

SNR Signal-to-Noise Ratio

SPL Sound Pressure Level

SVA Social Visual Attention

SVM Support Vector Machine

TIVA Task-independent Visual Attention

TOVA Task-oriented Visual Attention

VA Visual Attention

VAI Visual Attention Inference

WER Word Error Rate

WIMP Windows-Icons-Menus-Pointing Device

WIP Word Insertion Penalty

WoZ Wizard-of-Oz

WSJCAM0 Wall Street Journal Cambridge '0' corpus

CHAPTER 1

INTRODUCTION

1.1 Multimodal Interaction

Within the context of this thesis and multimodal interaction studies, the following terms and concepts are defined:

- *Modality:* A sense through which humans can receive computer output (i.e., output modality such as vision display, sound output) or through which the computer can receive human input (i.e., input modality such as speech, gaze, touch, gesture).
- *Multimodal Interaction:* The use of two or more modalities in an interaction.
- *Multimodal System:* A system that supports multimodal interaction.
- *Integration:* The process by which the information sensed in two or more modalities is combined/fused.
- *Recognition:* The process to detect and extract information from input modalities.
- *Information Event*: The occurrence of information in a modality represented by a sequential pattern of sensed data.
- *Relevance:* The usefulness of an information event from a modality (e.g., a gaze event) to the interaction or system purpose (e.g., a noise-robust speech recognition).

Conventional WIMP (windows-icons-menus-pointing device) is the prevalent implementation of graphical user interface (GUI). Disseminated by Microsoft Windows and Apple Macintosh, GUIs vastly change the way people interact with computers compared to the command-line system.

However, typically operated via keyboard and mouse, GUI limits the way people use the computer due to the lack of support for more natural and mobile interactions. Advances in hardware, software, framework, and algorithm are changing that by enabling significant shifts in the means people interact with computers. To accommodate a variety of tasks, scenarios, and users, systems must be more natural, adaptive, and perceptive. This becomes a primary motivation for developing multimodal systems.

While multimodal interaction happens between human-to-human and human-to-computer, the main focus in this thesis and computer engineering field is the latter. Systems that incorporate multiple modalities have been a shift away from conventional WIMP interfaces. The main objectives are to provide naturalness, flexibility, and real-world utility. Each modality has its relationships with other modalities and its own weakness or strength in terms of communication intelligibility and recognition difficulty. In a well-designed multimodal system, the use of multiple modalities should enable each modality to overcome another's weakness.

The information contained in the modalities may be complementary and essential to the realisation of system function [230]. Therefore, understanding and interpreting the behaviour of one modality may not be achieved without consideration of other modalities. The procedure to combine information from different modalities to achieve the better interpretation and interaction is called multimodal integration.

In addition to human speech, multimodal systems employ non-verbal modalities such as gaze, gesture, haptic and brain signals. The information sensed from the verbal and non-verbal modalities may be integrated to achieve specific system functions. The functions may range from the robust recognition of issued commands to making a machine more socially aware and 'human-like'. In a system that employs gaze and speech, the

examples of such functions can be using gaze to assist speech recognition [49] or to enable an artificial agent to convey impressions to users [89].

Within the context of multimodal research, this study considers the characteristics of the information recognised in modalities and their appropriate integration. One major contribution is the investigation of the selective use of information in one modality (eye gaze) to aid the recognition of the other (speech) considering real-world utility (environmental acoustic noise).

1.2 Automatic Speech Recognition

Speech is the primary means of communication between people. Research in speech recognition technology is motivated by the desire to give machines capabilities to understand and/or communicate with humans as humans do to one another.

Automatic speech recognition (ASR) is the process of a machine recognising human speech. ASR started in the 1950s and the development since has led to the emergence of the commercial speech recognition systems.

A standard ASR approach is based on the hidden Markov model (HMM) [260], which uses a probabilistic framework to estimate the most likely word sequences given the observed acoustic features. An acoustic model estimates the probability of the acoustic observations and a language model describes the probability of a word sequence [7].

Typically ASR systems evaluated in acoustically noisy environments degrades significantly compared to quiet laboratory assessments [56] [208]. Acoustic noise contaminates the speech signals and causes speakers to behave differently, leading to a mismatch between the speech used to train the system and that used in recognition. To overcome the problem, one typical approach is to reduce the mismatch with the help of the prior knowledge of the noise condition. Another approach is to introduce extra modalities. In this thesis, gaze is used as an extra input modality to integrate with speech and infer the level of acoustic noise.

The history of the ASR development, the effect of acoustic noise and the techniques to counter the noise (non-exhaustive) is reviewed in section 2.3.

1.3 Eye Tracking

Eye-tracking is a technology that allows a machine to be aware of humans' gaze movement behaviours. The advances in eye-tracking technology have spurred many psychology and physiology studies about the relationship between the use of gaze and the cognitive and perceptual processes [85] [317].

Following this, researchers started to investigate the use of gaze information in human-computer-interaction (HCI). The gaze information is used in the HCI studies to analyse human's interests and attentions [258], to replace mouse as a deliberate pointing modality [155] or to improve machines' capability in language comprehension [70] [301] and production [197] [104].

Eye tracking equipment can be broadly divided into two types - intrusive and remote. Intrusive devices need physical contact with the users, such as contact lenses, electrodes, and head-mounted devices. The head-mounted devices typically track the eye movement by measuring the light reflections from the eyeballs using head-mounted cameras [98]. The users' eye movements are recorded and decoded with image-processing techniques. However the attached equipment may cause discomfort or restrict users' natural head movements. Remote eye trackers are typically multiple cameras in the environment that fixate on the person's face. These eye trackers are believed to be easier to set up, but normally have less favourable accuracy. Compared to the remote equipment, the head-mounted eye trackers are more accurate (typically 2° [210] compared to 5° for remote ones [330]). In this work, a head-mounted eye tracker is used to ensure best possible accuracy.

Commonly, the outputs of an eye traker system involve fixation and saccade events. A fixation is an event when the eye is essentially stationary and a saccade is an event of rapid re-orienting eye movements between the fixations. The implementation in the

actual system often depends on the eye tracking equipment and software.

One major problem of using gaze in the interactional interfaces is the 'Midas touch' problem [137]; people involuntarily move their eyes and to use eye deliberately like a mouse consistently is against human nature. In the context of multimodal systems, because gaze is 'always on', the 'Midas touch' makes it reasonable to assume that not all information from gaze is useful for integrating with other modalities. Thus in this thesis, it is proposed that the 'relevance' of gaze be considered during the integration process with speech.

Discussions of gaze modality, the meaning of gaze in multimodal systems and automatic speech recognition are presented in section 2.2 and 2.4.3.

1.4 Gaze-contingent ASR

There is potential to use the information from non-acoustic sources to improve automatic speech recognition (ASR) performance. In multimodal ASR systems, the recognition performance can be improved using the information from non-verbal modalities such as gaze. Psycholinguistic studies have shown that eye gaze is highly related to human language processing and carries information about human attention [257].

Although information from gaze could improve ASR performance, earlier studies attest that the performance improvement for 'clean' speech (i.e., negligible background noise) is minimal [257] [49]; words not correctly recognised which are prone to speaker dis-fluency are not necessarily words semantically related to gaze events (i.e., nouns associated with visual objects); recognition errors have been demonstrated to be caused more likely by common short words, such as 'it', 'the', 'a', etc. [290]. It is reasonable to assume that the limited improvement space of recognising clean speech may be greater if ASR performance is compromised such as in acoustic noise. For this reason, it is reasonable to assume that the use of gaze in ASR is more beneficial in noisy environments. However, as far as the author is aware, no existing gaze-contingent ASR systems have been evaluated in such settings.

Environmental acoustic noise can greatly degrade ASR performance. Speech is harder to detect and there may be greater dis-fluency; a person tends to adjust their speech in noise for robust communication, leading to changes in spectral power, pitch and speech rate - the Lombard effect [146]. In this scenario, ASR could benefit from introducing gaze as an extra modality.

1.5 'Selective Use' of Events in Multimodal Integration

In a multimodal system, such as a gaze-contingent ASR, it is important to note that not all the events sensed contribute towards the system function and, hence, not all can be considered 'relevant'. Therefore, treating the relevant and irrelevant events equally during the integration process can compromise the outcome depending on the integration objectives. Thus, the use of events should be selective accordingly. In this thesis, these events are termed as 'task-oriented' and 'task-irrelevant' respectively.

To demonstrate the requirement to account for relevance and selective use of gaze events during multimodal integration, consider the following examples. In the seminal 1970s multimodal interaction system 'put-that-there' [283], a user positions shapes on a screen using speech and gesture. The user issues a command using speech that may be ambiguous ('put what where?'). Information to resolve ambiguities in speech is contained in the other modalities. In an updated version where gaze is sensed, the users could gaze at the location where they want to draw (a gaze event), orient their head towards it (a head pose event to assist in gaze and a natural response to centre one's vision), or describe the shape with their hands (an iconic gesture). Much of the work in achieving system function by integrating multiple modalities was undertaken in the late 1990s, e.g., Quickset [47]. However, in these systems, all observed behaviours (i.e., information events in modalities) were deemed relevant to system interaction. Move such interactive systems into a busy environment with a user who may be multitasking, and it becomes clear that

the multimodal recognition problem becomes less tractable. Speech may not be reliably sensed due to background noise. Not all gazes or gestures may be directed at the system - the users may be with other people doing other things, such as glancing at others or averting their gaze to concentrate. Thus, the user's interactive tasks assumed by the system may not be the only task being undertaken; the 'expected' multimodal behaviour may not be observed or mixed in with other irrelevant behaviours.

In this thesis, it is proposed that this relevance be considered as a prerequisite when using gaze information selectively in ASR.

1.6 'Selective' Gaze-contingent ASR

Information about what a person is looking at (or fixating upon) - their focus of visual attention - can be used in an ASR system to improve performance [49] [50] [257]. This is achieved by modifying word probabilities in the language model, a process called language model adaptation. This is possible because there is a semantic relationship between words and the foci of visual attention - i.e., names and attributes of objects and visual features.

A person's visual attention may not be related to their speech and not all gaze behaviour is related to visual attention; gaze behaviour can also be explained by cognitive process and interaction with others and the environment. Therefore to improve ASR performance using information from gaze, it should be used selectively.

In this work the meanings or roles of gaze are distinguished by their measurable validity. Those related to interaction or reaction to environment changes (e.g., system responses) can be measured and validated. The corresponding visual attentions can be used in language model adaptation to modify word probabilities.

On the other hand, the gaze behaviours related to cognition process cannot be easily validated. However, their relationship strength with speech can be estimated by assuming hypothesised cognition models, and it is proposed that the prevalence of different cognition roles changes as introducing different levels of acoustic noise. Thus, acoustic noise

condition can be inferred by exploiting this prevalence. The knowledge of noise can be used in acoustic model adaptation to reduce the mismatch between the ASR training and recognition conditions.

In short, the selective use of gaze in ASR is explored by acoustic model adaptation based on acoustic noise inference (Chapter 5) and language model adaptation based on visual attention type inference (Chapter 6). An application of selective gaze-contingent ASR is built and evaluated in Chapter 7. Related reviews are presented in Chapter 2 and more detailed system architecture will be described in Chapter 3.

1.7 Research Questions and Methodology

For the selective use of gaze in an acoustically noise-robust ASR, this research addresses the following questions:

- How to integrate the information events in gaze and speech considering their relationship (temporal and semantic)?
- Is gaze's behaviour and relationship with speech dependent upon acoustic noise? Can this dependency be exploited for ASR?
- How to use gaze selectively to integrate with speech by considering its relevance?

In the context of multimodal human-computer-interaction systems. The following methodology is followed:

- Development of a framework to model the relationship between the information events in gaze and speech.
- Collection of a corpus of eye movements and related speech recorded in acoustically noisy environments.
- Analysis of the dependency of speech, gaze, and their relationship upon acoustic noise condition and of the use of this dependency to infer the noise condition.

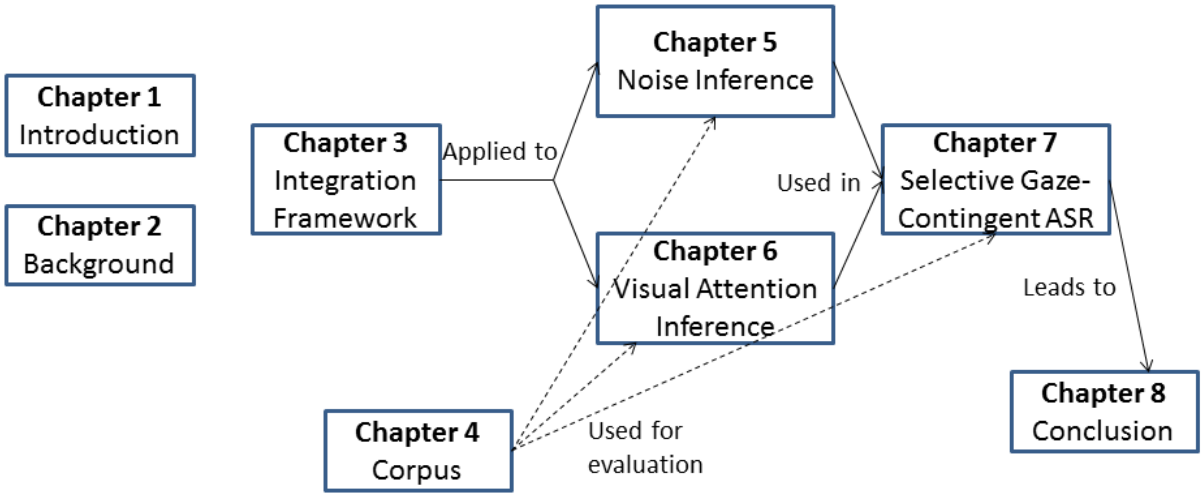


Figure 1.1: The hierarchical thesis structure. The meanings represented by the arrows are marked in the figure.

- Development of a framework to infer the visual attention type for the selective use of gaze in improving ASR performance.
- Construction of a baseline ASR system and a task-specific language model. Development of a framework to adapt the language model using the gaze selectively.
- Evaluation of the ASR system utilising the adapted language model and the inferred noise condition.

1.8 Thesis Structure

The thesis is constructed as shown in Figure 1.1:

- Chapter 1: An introduction to the background and the general objectives of the study.
- Chapter 2: A review of the work in related research fields, including gaze, speech, and multimodal systems.
- Chapter 3: The description of a general integration framework of gaze and speech for the acoustic noise and the visual attention inference in ASR.

- Chapter 4: A description of an eye/speech corpus (ES-N) collection in different clean and acoustically noisy environments.

- Chapter 5: An analysis of the behaviour changes in acoustic noise for speech (acoustic Lombard effect) , gaze (which is termed as 'gaze Lombard effect') and their relationship. A description of an information-theoretic-based measurement for the speech-gaze relationship and its use in the acoustic noise inference.

- Chapter 6: A description of a visual attention type inference framework based on the relevance in integration with speech for noise-robust ASR.

- Chapter 7: A description of the construction of a research-level baseline ASR, a task-specific language model, and the language model adaptation approach using the gaze information selectively. A demonstration of incorporating acoustic noise inference in ASR by acoustic model adaptation. An evaluation of the ASR performance based on the findings from Chapter 5 and Chapter 6.

- Chapter 8: A conclusion of the contributions and the recommendations for potential future researches.

CHAPTER 2

GAZE, SPEECH, AND MULTIMODAL SYSTEMS

2.1 Multimodal Interaction

2.1.1 Multimodal systems

A multimodal system combines information contained in input modalities, such as pen, touch, speech, deictic gestures, eye gaze, and head and body movements in a systemic manner to fulfil a system function. It produces multimedia output, such as sounds and screen displays. In the context of human-computer-interaction (HCI), the evolution of these systems is driven by new input and/or output technologies. One of the first multimodal systems, the seminal 'put-that-there' [283] developed in the late 1970s, combines speech and hand-pointing with technologies available at the time: magnetic field sensing via inductors to detect hand pointing. This system has different technologies than those used today; hand pointing is more likely to be achieved by vision-based methods, such as the Kinect sensor [223] or accelerometers [329]. Other early systems combine speech and mouse pointing, such as the CUBRICON system [218], or recognise commands via speech while determining the pointing target from gaze or manual gestures, such as ICONIC [167]. Despite the differences in technologies, the lessons learnt about how people interact multimodally are still relevant today.

A key aim for multimodal interaction in HCI is for the computer to communicate with

users in a more naturalistic (i.e., human) manner. Recent multimodal systems are capable of recognising a broader range of input modalities with newer technologies. Our voice, hands, gazes, and gestures (hand and body) can be tracked by different sensors such as microphones, depth cameras, and accelerometers.

For the foundation of multimodal system design, empirical results have led to heuristic design principles. For example, Oviatt lists 'Ten Myths of Multimodal Interaction' [230], offering insights such as, 'Users do not always interact multimodally', and 'Multimodal signals do not always co-occur temporally', which enlightens the researchers in the field. A later survey by Jaimes [139] discusses major vision approaches for multimodal interaction involving gaze detection, gesture recognition, and body movement. It highlights that the context information (affect) is an important aspect to consider because it may influence the behaviour of humans, such as facial expressions, gestures, and tone of voice.

The progress in both hardware and software enables more modalities to be used and improves the utilisation of existing ones. For example, the early development in gaze-tracking devices allows the use of eye movement to be a potential measure in controlled laboratory environments [324]. Many traditional gaze-tracking devices require physical contact with the user, such as contact lenses, electrodes, and head-mounted devices. The gaze-tracking techniques and various studies [137] have enabled gaze to become a popular modality in HCI multimodal systems. However, the attached gaze-tracking equipment can cause discomfort or restrict users' natural head movements. Later, progress has been made to develop remote gaze-trackers which are claimed to offer more comfortable use and faster setup [210]. However, limitations, such as unsatisfactory accuracy and stability, still exist in the current remote gaze-tracking technologies that continue to motivate the studies in more robust hardware or software solutions. For example, novel gaze-tracking techniques allowing less restricted head movement and faster calibration are reported [331]. In a state-of-art review [107], gaze-tracking systems is presented from the different methods of detecting eye images to computational estimation models and gaze-based applications.

The development of the system framework and algorithm allows the better design of

multimodal systems. Dumas [69] presented a survey of principles, models, and frameworks in designing multimodal systems. Not exhaustively, the survey covered many hot topics in theoretical principles, time-sensitive software architectures and multimodal integration. As a consequence of the growing studies in basic architectures and frameworks, real applications are built. These applications range from map-based interaction systems [49] and medical applications [187] to virtual reality interactional agents [128] and so on.

2.1.2 Use of modalities

In early multimodal systems, although extra modalities are used in addition to the speech, their usage is limited. While instructions are given via speech in the 'put-that-there' system, the deictic term (such as 'there') is conveyed by manual pointing. In other systems, this pointing information may be recognised via pen [233] input. Compared to speech-only systems, the extra modality in these early multimodal systems only serves as the function of a mouse with worse performance. For example, the accuracy of pointing using hand gestures can range from below 50% to 90% depending on the user [284].

Users do not always interact multimodally (i.e., make use of extra modality). In one study, it is reported that the speak-and-point patterns only bear 14% spontaneous multimodal inputs [233]. Also in the systems where gesture served as an input modality, the pointing gesture is stated to account for only 20% among all gestures [57]. However, modalities that convey writing and drawing, body gestures, and facial expressions can provide richer multimodal information in addition to pointing. For example, pen input is used more frequently for drawing symbols, signs, or digits, and gaze can be used to manage conversation and user focus.

Later studies have investigated a variety of ways to utilise multimodal inputs. For example, lip movements are combined with speech for better recognition [40] [135] and video information such as position, velocity, and size are combined with audio information for speaker tracking [22] [300].

2.1.3 Robustness of recognition

Major motivations of using multiple modalities include improving the robustness of the information recognition in modalities and combining them to fulfil a system purpose. A well-designed system that uses multiple modalities can achieve higher stability/accuracy by using information in modalities and the information relationship between modalities to reduce errors in each single modality.

There was previously concern regarding combining two error-prone modalities as it had potential to compound errors and harm the reliability of the system [230]. However, cumulative studies have claimed that combining two or more modalities is an effective way of removing recognition uncertainty. For example, improved recognition rates have been achieved by combining speech with spoken animation [211], lip image sequence [214], side-face images [135], face-detection features [219], pen-based gestures [234], and gaze movements [49] respectively.

Using multiple modalities can make a multimodal system support better error-handling. When users are allowed to interact with a system multimodally, they tend to simplify their languages, which reduces the complexity of speech processing. For example, when completing system tasks (typically map-based) that involve spatial description, users involuntarily prefer less complicated spoken instructions with the help of pointing modalities. Oviatt [235] compares the linguistic structure where the users interact in speech only and multimodally with a pen in a spatial map task. It is reported that when they interact multimodally, the users' language is less complex in terms of fewer referential expressions, determiners, and noun phrases. In a recent study [75], a 'put-that-there' task is conducted in near (0.5m) and far (2m) distances employing speech and hand pointing. The study reports that out of 80 times in both distances, shorter instructions as 'that there' and 'select drop' are chosen 76 times while the longer instruction 'put that there' is only used 4 times in near distance.

Another important factor that boosts error-handling capability is the users' freedom to select their input modality. Users tend to select the less error-prone modality when they

are free to choose. For example, users are more likely to write a foreign surname rather than speak it [236]. Moreover, users tend to switch input modality following a recognition error, which can be an effective shortcut to avoid repeated failures [231]. Last but not least, an interesting finding from a usability test [231] claims less subjective frustrations with the system errors when users are free to interact multimodally, even when the error rate is as frequent. It is suggested that the phenomenon is related to the greater sense of control brought by the freedom.

2.1.4 Context awareness

In order to better integrate the modalities to improve the quality of interaction with humans, a desirable design objective is to increase the amount of information available to the system. This information gain is called context awareness (CA) [1]. Factors that contribute to CA include background (domain) knowledge and knowledge pertaining to the physical environment (e.g., acoustic noise condition), which may be utilised to improve system performance. In systems employing gaze as an input modality, CA can be utilised to assist the interpretation of a gaze event and measurement of the relationship to the system task.

The relevance or irrelevance of non-verbal modalities (e.g., gaze events) to a social interaction is highlighted by a distinction between what is expressed (relevant) and what is experienced (irrelevant) [221]. For example, what is expressed is what is acted, e.g., gaze events that contribute to an interaction. In contrast, what is experienced is a person's cognition, e.g., gaze events irrelevant to an interaction that convey no social purpose. However, from an engineering perspective, when interpreting or inferring a non-verbal modality such as gaze, the observer can infer both experiential and/or interactional meaning. This creates the engineering problem of inferring the interactional meaning and discounting events related to experiential meaning. It is a common problem to other non-verbal modes of communication, such as gestures, and has long been recognised in cognitive psychology. In one of the first taxonomies of non-verbal behaviour such expe-

riential meanings are classified as cognitive states [74]. Systems that attempt to infer cognitive states refer to them as 'cognitive context' [34].

In this study, a user's communication intent is defined as what he/she tends to express during a system task. This intent is also known as the 'interactional intent' or 'interactional meaning' behind a person's verbal and non-verbal multimodal behaviour [221]. To infer the communication intent of a user, it is necessary to recognise, as events, those patterns in non-verbal behaviour that contribute towards the interaction and to measure their strength of relevance to the interaction. Accordingly, irrelevant events must either be explicitly identified by an inference model or implicitly modelled as variables containing a random element, e.g., stochastic noise. For example, speech recognition systems commonly employ a 'babble model' of real-world non-stationary noise representing irrelevant sounds and background speakers [308].

2.2 Gaze in Multimodal Systems

2.2.1 Selective use of gaze

Gaze has been exploited as an input modality in human-computer-interaction (HCI) systems. Early systems, such as the 'put-that-there' interaction system [283], combining speech and gesture were followed by systems that combine speech, gaze, and gestures [167] as well as speech and writing [47]. More recent gaze-based systems also incorporate information from a person's emotive state [245], brain computer interface [199], mobile devices [313], and multitouch devices [120].

In 'traditional' HCI systems, gaze is originally used as an active modality where a user deliberately uses it to initiate or convey a command. For example, gaze can be deliberately used to point and select menu items, replacing the function of a mouse [131] [220] [138] or type on a graphical keyboard [280] [189].

On the other hand, there has been a shift towards developing 'attentive' or 'perceptual'

user interfaces [304] that assume more natural gaze behaviour without requiring the user to deliberately use it. Compared to 'traditional' interfaces, information from a user's gaze is utilised in a manner closer to how it is used when interacting with another person. This requires the machine to sense multiple modalities during an interaction and recognise social behaviours accordingly [314].

The 'Midas touch' problem [138], where a person involuntarily issues a command to a system because gaze is 'always on', has motivated the development of techniques that overcome this problem. The examples of these techniques include magnifying portions of a screen [8] and the use of speech to disambiguate words [328]. A specific gaze event such as a fixation upon an object will have particular characteristics - for example, the dwell time, fixation duration, pupil size, and the length of saccade towards and from the object. The 'Midas touch' problem can be seen as a problem of how to selectively use the gaze events relevant to the system function and discount the irrelevant ones. In the 'traditional' HCI interfaces, the selection can be realised by setting dwell time threshold [138] [242] or using eye blinks and winks as signs [287]. In 'perceptual' interfaces where more natural interaction style is allowed, the selective use of gaze can be realised by analysing gaze characteristics. For example, in the case of reading detection, gaze is selected based on the saccade length and movement direction [37], and for several different tasks, the gaze sequence is selected by tracing eye-movement protocols [281]. It is also shown that whether a gaze event is related to a task-relevant command can be automatically learnt using the fixation, saccade, and pupil features [21].

It needs to be noted that in some multimodal systems, gaze is recorded for attention monitoring. In these systems, where other modalities, such as speech, are used to complete system function, gaze is recorded and analysed to check user understanding or modulate the content provided based on users' interest [13] [188] [213]. In these interfaces, gaze is not used for direct interactions with the system. Thus, the use of gaze in a gaze-contingent ASR can be considered attention-monitoring.

2.2.2 Gaze features

To interpret the gaze data, some gaze features in the data stream must be extracted. Eye-trackers produce gaze signals including eye orientations and/or position of points on a display object (e.g. a screen). One of the first steps in analysing gaze data is to extract fixations (times when the eye is essentially stationary) and in-between saccades (rapid re-orienting eye movements) [137].

Fixation duration and saccade length are features most commonly reported in gaze studies [137]. Fixation duration is normally believed to be correlated with the difficulty extracting information, mental workload, and internal processing [87] [99]. For example, a longer mean fixation duration is shown to be related with raised attention and processing depth during problem solving [137], and web browsing [72], or higher information priority in reading [68]. The mean fixation duration is analysed in gaze studies such as reading text on a screen in various formats [166], selecting items from computer menus in various styles [111], symbol search and counting on colour or monochrome displays [11], analysing web search and judgement [76], and extract information from web pages [53] [100].

Similarly, saccade length is also believed to be sensitive to workload (i.e., related to task difficulty) [30]. While not used as widely as fixation duration, saccade length is reported in gaze studies such as selecting command button specified from buttons grouped with various strategies [168], and extract information from web pages [100].

During the interaction with a system, not all gaze events are contributing to the system task (i.e., task-oriented, see section 1.5 and 2.2.1) thus termed as task-irrelevant (e.g., confusion, looking away, etc.). Based on the finding that the task difficulty influences both fixation and saccade features [248], Rayner [263] discusses the complexity of the correlation between fixation duration and saccade length. He stated that in eye-reading research, during reading (task-oriented) situations, there is no correlation between fixation duration and saccade length [267]. Whereas there is a correlation in non-reading (task-irrelevant) situations, the longer the saccade is, the longer the next fixation [216] [153].

Pupillary response such as pupil size change is another widely investigated gaze feature

and considered related to the mental workload. Pupillary responses are shown to be task-dependent and corresponding to the task difficulty [17]. Based on the finding, the pupil size is used in a framework to detect the task boundaries [12] [132], or to distinguish the situation between normal reading and looking for a given information in the case of a map reading and searching task [163].

Pupillary responses can be triggered by internal mental processes (e.g., attention, affect, mental workload, etc.) and external states (e.g., touch, visual and audio stimuli, such as acoustic noise) [149]. Although the pupil size change is correlated with cognitive intensity and considered reflection of internal states, the change caused by external stimuli is distinctly larger [19]. Using fixation and saccade features to predict users' intention of issuing a command in an HCI system is proved to have great potential with an accuracy of 75.1%, and this accuracy is slightly improved by 0.8% by adopting pupillary response as an extra feature [21].

2.2.3 Visual attention

Jakob [137] presented a review of the gaze features used in the HCI studies and he anticipated that learning user's deployment of visual attention leads to interfaces more closely fit to human needs. Visual attention (VA) is a specific type of 'gaze information event', which is closely related to the psycholinguistic processes (e.g., speech production) [49]. Thus, VA information produced by the eye-trackers is an important cue for a gaze-contingent system. Cooke [49] shows that the current focus of visual attention indicates an increased chance of a person's speech related to this focus, and thus, can be exploited in a gaze-contingent ASR.

A definition of 'Attending visual foci of interest', or more succinctly, 'Visual Attention' is provided by Harris and Jenkin [110]:

- 'Attention implies allocating resources, perceptual or cognitive, to some things at the expense of not allocating them to something else'.

VA is closely coupled to (although not the same as) a person's fixation direction [84] [63]. VA has been modelled from the perspective of human and computer, with the former's theoretical models influencing latter's applied models. In gaze-related HCI studies, the analysis of VA can be either *top-down* - based on factors such as user cognitions and goals, or *bottom-up* - based entirely on observation of interaction patterns without trying to infer the cognitive activity [100] [137].

The top-down and bottom-up factors can be distinguished by their measurable validity. A top-down approach may seem attractive because it involves inferring cognitive processes from gaze data. However, the top-down factors which affect visual attention are less easy to successfully engineer due to the lack of measurable validity. In addition, researchers do not always have strong hypothesis or theories to drive the analysis. On the other hand, the techniques for modelling bottom-up visual attention (e.g. the 'salience' of speech pattern [256]) are better understood as they concern objective measures of the interaction, thus can be validated and more easily engineered. Even when theories are available to support the investigation, a bottom-up approach can be rewarding if the gaze data and external stimulus (e.g., system responses in an interactive interface) are analysed properly [137]. For example, VA foci are investigated to analyse users' eye movement patterns in the case of consumer choice [276], or daily activities, such as making a cup of tea [172].

There are many factors which may contribute to a person's VA and related gaze behaviours, and modelling these remains an active topic. In this thesis, both cognition-driven top-down and interaction-driven bottom-up approaches are investigated and different VA type is inferred for the selective use of gaze in ASR systems. These two approaches distinguish the meaning of gaze in terms of measurable validity during HCI interactions.

2.2.4 The meaning or role of gaze

Observed gaze behaviours can be explained by a variety of meaning or roles. For example, gaze behaviour has been ascribed several roles relating to a person's cognition and its role in interaction [105]. Because gaze has different meanings or roles during interaction, the

utilisation of gaze in interactive systems may benefit from the selective use of gaze based on the gaze role inferred or assumed. For example, in ASR systems, assumptions of psycholinguistic processes have been exploited in statistical language model adaptation to improve the recognition performance [257] [49] [50] and to resolve referential ambiguity in speech when people navigate virtual worlds [253]. For interactive systems to infer the gaze roles for selective use in the modalities integration, the corresponding cognitive and interactional processes that manifest gaze behaviours must be understood.

The fields of cognitive and social psychology provide insights that may be potentially exploited in inferring the meaning of gaze. Early pioneering psychological studies linked fixation durations to cognitive processes (i.e., perceptual processes) [322] and social roles (e.g., modulating social interaction) [157]. Later, the technological advances in camera-vision-based eye-tracking hardware and eye-movement detection algorithms (see a technology review by Hansen [107]) allow studies affording the measurement of finer granularity of the task-specific durational and spatial characteristics of gaze. Task-specific fixation duration and saccade length distributions have supported the development of cognitive models, such as scene perception, reading, object naming [266] [264] [265], and lexical processing [104].

Gaze roles related to cognition can be individually elicited in controlled psychology research experiments (e.g., application of eye tracking in speech production [198]). However, in less constrained conditions, eliciting and attributing gaze behaviours to specific cognitive processes are problematic because of the absence of measurable validity. For example, one could assume that someone glancing around an environment could be committing the scene to memory while his/her cognition may be entirely unrelated. In addition, a specific cognitive process may not correspond to the onset and offset of an identified gaze event (e.g., a fixation event).

Although models of cognition from psychology have been adapted or used in interactive systems to give human capabilities to machines (e.g., see the review by Vernon [311]), the hypothesized cognitive models are less relevant to the completion of system function due

to their emphasis on understanding brain function rather than interaction with machines. For example, when people interact, their visual attention influences one another, and the consequent 'joint attention' increases communication effectiveness [269] [270]. This has led to the development of a system where the use of two eye trackers allowed people to see one another's attention on the screen [39], and to the efforts in replicating this behaviour when communicating with robots [293].

The distinction of gaze roles in terms of measurable validity is related to the distinction in multimodal behaviours stated by Norris [221] in terms of what is expressed (relevant) and what is experienced (irrelevant) (see section 2.1.4). In contrast to relating gaze behaviour to cognition (cognitive context awareness CCA), it can be more reliably interpreted in relation to a system task or activity (interactional context awareness ICA), communication with others, and reaction to changes in the environment (environmental context awareness ECA) [34]. In such instances gaze can be considered to have an 'interactional meaning' [221], and unlike cognition-oriented roles, validity is measurable due to observable variables and established taxonomies. For example, it is demonstrated that the gaze events related to instructing a system to move a block on the screen can be labelled and inferred with the observed gaze features of fixation duration, saccade length and pupil size [21].

In this thesis, a working taxonomy for gaze roles is proposed for the purpose of using gaze selectively to improve the noise-robustness of the ASR system. The taxonomy and an integration framework will be further discussed in Chapter 3

2.2.5 Machine learning approaches in gaze analysis

Gaze studies face a central problem that is the insufficient understanding of the link between the low-level gaze signals, measurements, and high-level behaviours and interactional events. Machine learning (ML) and classification techniques provided a working means of investigating and exploiting this link.

ML approaches have been adopted in gaze studies to process gaze data in great vol-

umes automatically. For example, a hidden Markov model is applied to uncover the processing state of a user in information searching tasks using eye movements [291], and achieved above 60% accuracy. ML with gaze data has also been applied in biometric person authentication [154] [160]. With gaze data such as eye movement velocity, pupil size, and gaze direction, the ML approach achieved an identification rate of 60% [20].

In the case of inferring cognitive states during problem-solving [73], and inferring gaze events related to issuing a command during an HCI interface [21], the Support Vector Machine based approach is applied employing combined gaze features (e.g., fixation, saccade, pupil size), and the accuracy of 53% and 73% are achieved respectively. The issues of exploring appropriate method and feature selection for analysing gaze data using ML approaches have been investigated in the case of revealing user interest during search or reading [2], and inferring object relevance in dynamic virtual scenes [152].

ML approaches have been demonstrated capable of inferring cognition or behaviour patterns from the low-level gaze data. However, finding the efficient techniques for classification and feature engineering still remains an active topic and needs to be explored. In this thesis, ML approaches are adapted to infer the noise condition (will be discussed in Chapter 5) and visual attention type (Chapter 6) based on the observed gaze features and the coupling with speech.

2.2.6 Gaze and speech in multimodal systems

Although gaze is not directly bound to speech production in terms of the articulatory process, a close correlation has been found between speech and gaze in the context of psycholinguistics. Gaze is involved in both language comprehension and language production process. For gaze in language comprehension, experiments are performed where the participants are presented with displayed objects while listening to the speech describing these objects [70] [301]. A temporal relationship is shown to exist between the eye movements to the objects and the spoken words referring to these objects; eye movements followed spoken words by $250ms$ in average.

For language production, gaze direction has been found to be related to spoken words when describing visual scenes. In a study [197] where the speakers are asked to name the objects on the screen with their gaze movement tracked, the results suggest that the speakers fixated on an object with a mean duration ranging from $740ms$ to $805ms$ before naming it due to the linguistic planning processes. This phenomenon in object-naming is also investigated in other studies, and the average latency between fixating an object and mentioning it is reported to be $902ms$ [104] and $932ms$ [105] for a mean duration of $600ms$. This latency is referred as the 'eye/voice' span - time between the onset of the gaze to an object and the subsequent onset of the spoken word referring to that object.

The complementary information that exists in gaze for speech production (i.e., object names) has motivated studies in gaze-speech multimodal systems. Because gaze fixations are related to the spoken objects in the scene, many systems have explored the use of gaze as a deictic (i.e., pointing) modality. For example, a dialogue system is proposed by NASA (National Aeronautics and Space Administration) for an assistant robot to infer what a user is referring to using gaze information [36]. For instance, if a user looks at the crew hatch just before saying 'door' in the command 'open that door', the robot would infer the command as 'open the crew hatch door'. However, the system performance is not reported. Another study [155] investigates the integration of speech and deictic information in gaze for a 'move-it-there' task. When a user issues an instruction to move an object on the screen by saying 'move it there', the object to be moved is selected based on what is being fixated on. These systems assume that a multimodal ASR utilising gaze (i.e., a 'gaze-contingent' ASR) is error-free.

2.3 Acoustic Noise and Its Impact on ASR Performance

2.3.1 Automatic speech recognition

Research into speech technology started in the 1930s at Bell laboratories. One of the earliest automatic speech recognition (ASR) techniques is reported in a Bell Labs paper in 1952 [60] with the capability to recognise isolated digits of a single speaker. The system relies on the formant frequencies estimated during vowel regions of each digit. Other laboratories involved in the ASR researches in the 1950s include RCA Lab, which reported a system that recognised 10 syllables of a single speaker [224], and MIT Lincoln Lab, which built a speaker-independent 10-vowel recogniser [88].

The technology advances in the 1960s include the concept of speech segmenter, which was first used by Sakai and Doshita from Kyoto University [277], and the concept of 'non-uniform time scale', which was demonstrated through the work of Martin [192] and Vintsyuk [315]. The use of dynamic programming methods between two utterances for similarity assessment (generally known as dynamic time wraping DTW) proposed by Vintsyuk is earlier than the more formal methods reported by Sakoe and Chiba [279]. Despite this, the fact that dynamic programming, including the Viterbi algorithm [316] has become an indispensable algorithm in ASR since the late 1970s is mainly due to the superior performance published by the latter.

In the early 1970s, the concept of linear predictive coding (LPC) simplified the speech decoding from the large data requirement [9] [133]. The technique was later applied to the ASR systems proposed by Itakura [134], Rabiner [259], and others. Another milestone is the effort from IBM in large vocabulary speech recognition. The speaker-dependent system, Tangora [141], focuses on the recognition vocabulary size and the probabilistic structure of the statistical language model (LM). Being a variant, the *n-gram* model, which characterises the probabilistic relationship of the contiguous n words, is used most frequently in current ASR systems. On the other hand, effort is put by AT&T Bell Lab

into a speaker-independent system [259] that can deal with many different speakers with various accents. In addition, the concept of *keyword spotting* is later proposed by the lab as an evaluation metric for ASR systems [320].

In parallel to the efforts from these two labs, a large-scale speech understanding project is funded by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defence. The project spurred many seminal studies and systems, including the Harpy system [186], the Hearsay-II system [79] from Carnegie Mellon University (CMU), and the Hear What I Mean (HWIM) system from Bolt Beranek and Newman Inc. (BBN) [162].

From the 1980s, the focus of the ASR research became the recognition of continuous words. There was a technology shift from the template-based approaches to statistical modelling frameworks, with the didden Markov model (HMM) being the most notable methodology [178] [260]. HMM is a probabilistic model with an underlying stochastic process that is not observable (i.e., hidden) but that can be estimated from a sequence of observation. The basic concept of HMM is early known by some laboratories, such as IBM and the Institute for Defense Analyses (IDA) [83]. However, it is after the mid-1980s that the framework was completed [178] and became a widely applied methodology for ASR. The widespread use of the HMM-based frameworks has continued since due to the continual improvements and refinements to the algorithms. More reviews for the applications of HMM in ASR can be found in the publications of Gales [93] and Oviatt [226].

Due to the great success in statistical modelling and the interests from ARPA, some novel systems were developed in the early 1990s, including BBN's BYBLOS system [44], SRI Group's DECIPHER system [212], and CMU's SPHINX system [174]. The SPHINX system employs discrete HMM, which uses discrete distributions to model acoustic vectors. The SPHINX system can perform speaker-independent continuous speech recognition in real time by applying pure statistical methods and achieved more favourable results compared to the other two systems.

Another milestone in the 1990s is the speech recognition tools that spurred many individual speech recognition studies. One most successful and widespread tool is the Hidden Markov Model Tool Kit (HTK) developed by Cambridge University [325]. The HTK enables the set up of a well-structured baseline speech recogniser that allows the contributions in new concepts and techniques to be made from a global community.

Another notable technology is the artificial neural network (ANN). First introduced in the 1950s with less satisfactory results [195], ANN is re-applied to the ASR systems in the late 1980s. The model was initially used in some simple recognition tasks, such as recognising a few phonemes and words [182]. ANN was more difficult to train - particularly the large/deep networks for ASR. It is worth mentioning that, due to the developing concept of the *deep belief networks* [118], the recent use of ANN in ASR has again gained increased attention for acoustic modelling [117], phoneme recognition [205], and large-vocabulary speech recognition [54].

From the 2000s, in addition to the continual evolvement of the statistical models and deep learning techniques for the ASR, there is a shift in the research focus to the ASR evaluation in acoustic noisy environments (noise-robust ASR) and performance improvement by introducing extra modalities (Multimodal ASR). These aspects will be discussed respectively in section 2.3 and section 2.4. The study in this thesis addresses the noise-robustness of the ASR by adopting gaze selectively in a multimodal framework.

2.3.2 Effect of noise in ASR performance

It is typical that recognition rates in laboratory assessments do not necessarily represent the performance in natural environment settings due to the noise, interruption, increased cognitive load, and human performance errors in the natural environments. Researchers started to investigate the performance in natural environment settings by evaluating the systems on noise-corrupted data. For example, an early study [208] evaluated the speech recognition accuracy on a corrupted Wall Street Journal corpus. The clean speech was artificially injected with additive noise, and an accuracy drop up to 60% was reported. In

an experiment conducted by IBM [56], a 99%-accuracy recogniser was reported to have a more than 50% recognition rate drop in a cafeteria environment.

There are two broad types of noise. 'Stationary' noises (e.g., white noise, road noise in a moving vehicle, and so on) are relatively easy to model and process, as they are more predictable. However, 'non-stationary' noises are more common in natural environments. They either change irregularly (e.g., multi-speaker babble noise) and/or involve variable phase-in/phase-out noise (e.g., due to speaker movement). Thus, they cannot always be predicted or modelled.

The performance degradation under environmental noise has been viewed as a primary obstacle to the commercial use of speech recognition technology [101] [146]. There are two main problems caused by the acoustic noise in the environment which can elicit considerable mismatch between the training and recognition conditions, resulting in serious degradation in recognition accuracy. First, the noise itself contaminates the signal vectors representing the speech, resulting in increased processing difficulty and, secondly, the acoustic Lombard effect [146].

2.3.3 Acoustic Lombard effect

Acoustic noise has been demonstrated to change people's speech behaviour. The change is known as the Lombard effect [146]. Although the nature of the Lombard effect is speaker-dependent, some common changes in terms of spectral power, pitch and speech rate are reported [127]. The magnitude of these changes is shown to be dependent on the speakers' desire for intelligible communication [106] [71]. The Lombard effect in an interaction between people is likely to be more significant than where speakers are reading a list to themselves: 'The speaker does not change his speech behaviours to communicate better with himself, but rather with others' [173].

The estimation of speech intelligibility in noise is an important metric to consider when measuring the Lombard effect [148]. Speech parameters, such as vocabulary size [121], word duration, and vocal effect [246] [66] [307] vary in acoustic noise to aid communication.

In the context of multimodal interactive systems, where the user is expected to use his gaze to aid speech in noise, the dependency between gaze and speech can be argued to be a metric of communication intelligibility. Thus, it is of interest to explore this metric.

Lombard speech differs between people and noise type [206] [147]. Two commonly used noise types for comparison are stationary white noise and multi-talker babble noise. Babble noise introduces stronger speech changes compared to white noise [147] [148]. A recent study [204] shows that babble noise is a stronger masking noise for affecting speech changes and introducing higher gaze frequency.

2.3.4 Strategies for noise-robust ASR

A typical approach to improve noise-robustness in speech recognition systems is to reduce or remove the mismatch between training and recognition data or environment. The basic idea of the approach is to either transform the speech data to match the data recorded in training environment, or transform the model parameters trained to better match the noisy environment. To realise noise-robustness, the approach can be broadly divided into three groups of techniques:

1. *Speech enhancement:* The noisy speech data is transformed to a reference condition (normally clean-speech condition) to resemble the speech recorded in that condition, and recognise it using the system trained in the reference condition [101] [59].

2. *Feature enhancement or model compensation:* These techniques try to decrease the mismatch between the model trained in the reference condition and the data observation by transforming the model features/parameters to better resemble the noisy speech distribution [101].

3. *Noise-resistant feature extraction:* This technique assumes a recogniser is noise-independent. The noise-robustness is realised by extracting noise-resistant (i.e., less distorted by noise inherently) speech features and exploring robust distance measures [101] [142].

Most speech enhancement techniques assume noise is additive and stationary over a relatively large time window. Thus, their performances are less satisfactory in dealing with non-stationary noise. Speech enhancement is realised by transforming the noisy speech data; therefore, distortion is inevitably brought to the speech itself. Some speech enhancement techniques were originally developed for speech quality improvement rather than speech recognition. While these techniques can be used as a pre-processing step before recognition, each does not necessarily improve the recognition accuracy because the introduced distortion can be tolerable to human listeners but not recognisers. Speech enhancement algorithms improve speech quality but not necessarily speech intelligibility [126] [125]. Loizou [185] suggests that the causes are the perceptual effects of the distortions on the intelligibility, and the fact that none of the techniques were designed to maximise the intelligibility metrics. Some examples of the speech enhancement techniques are subspace-based methods [78] [116], spectral subtraction [24] [297] [225], statistical-model-based [77] [237], Wiener filtering [310] [179], and the vector Taylor series (VTS) compensation algorithm [209] [150]. Hu [124] reported a subjective comparison performed on spectral subtractive, subspace, statistical-model based, and Wiener algorithms. It is believed that, in general, the statistical-model based methods performs the best, followed by the multi-band spectral subtraction method [151].

For the model compensation techniques, rather than processing the noisy speech to remove the noise corruption, the parameters of recognition model are adapted to account for the presence of noise. For example, the HMM [260] provides a framework to model temporal and spectral characteristics in speech signals. HMM can be trained to model the clean speech specifically and then be adapted to the noisy speech by changing parameters, such as mean and variance of a Gaussian distribution. The model compensation technique can potentially allow the optimisation of the model parameters to compensate for the noisy conditions not presented in the training stage. As the adaptation relies on the noise conditions the compensated model cannot be generalised to deal with all different noises. Some examples of model adaptation methods other than the HMM decomposition

technique [309] include maximum-likelihood linear regression (MLLR) [175], maximum a posterior probability (MAP) [97], and parallel model combination (PMC) [94].

For the noise-resistant feature extraction techniques, focus is laid on extracting the speech features that are less distorted in noise environments. By employing the noise resistant features in speech recognition, recognisers are likely to be less sensitive to the effect of noise; therefore, recognition can be potentially seen as noise-independent. To compare the feature in different environments, similarity measurements are also studied. While the speech enhancement and model compensation techniques require prior knowledge of the noise, one advantage of the noise resistant feature extraction techniques is that weak or no assumptions of the noise are made. On the other hand, this can be a shortcoming as the performance cannot be optimised without making use of the robust features according to a specific noise condition. There has been a variety of noise-resistant feature extraction methods attempting to derive features that are more consistent under noise; examples include root-cepstrum coefficients (RCC) [3], the modulation spectrograms [159], spectral peaks [14], perceptual linear prediction (PLP) coefficients [114] combined with the relative spectra (RASTA) techniques [115], and the worth-mentioning Mel-frequency cepstral coefficients (MFCCs) [61], which have become the standard feature set in ASR.

In contrast to the recognisers, human listeners are more capable of dealing with contaminated speech by utilising the partial information and reliable regions left in the distorted signals. Inspired by this phenomenon, missing feature theory (MFT) [48] is proposed assuming some of the time-frequency regions of signals are dominated by speech, while others dominated by noise and speech recognition can be performed on the more reliable components. One advantage of MFT is that its similarity to the human auditory system in dealing with partial information in noisy data makes a minimum assumption on the noise condition. Instead, reliable regions are identified, and the recognition is performed accordingly, making the recogniser effective in compensating either stationary or non-stationary noise. A study for the effectiveness of MFT was reported by Raj [261], in which a wide variety of MFT techniques and masking schemes were reviewed. A more

recent study [306] confirmed the robustness of MFT in large-vocabulary ASR.

Although the techniques described above claim to improve speech recognition performance in noisy environments, the optimum performances reported are more likely to be achieved in controlled noise conditions. One reason is that these systems are commonly evaluated under stationary noise. While stationary noise (e.g., white noise) can be modelled and processed more easily, many noises in natural environments are non-stationary (e.g., babble noise in commercial areas). Although techniques like noise-resistant feature extraction or missing feature approaches claim to make less assumption on the noise type, their performances when dealing with non-stationary noise degrade significantly compared to when dealing with stationary noise [261].

These strategies are concernd with the acoustic components of the ASR systems (e.g., an acoustic model) and not the linguistic components (e.g., a language model). In this thesis, the selective use of gaze in ASR is realised by both acoustic and language model adaptation (presented in Chapter 7).

2.3.5 Lombard effect and ASR

A study in the 1980s [262] first demonstrated that the Lombard effect can degrade the speech recognition more than the noise contamination itself. Also Junqua [145] reports that the Lombard effect is more degrading than the additive noise for a speaker-dependent recogniser as it assumes minimum intra-speaker variability.

While the additive noise contamination can be better dealt with using a variety of techniques (see section 2.3.4), the mismatch between training and practice introduced by the Lombard effect is more difficult to deal with. The difficulty in processing speech under the Lombard effect (i.e., Lombard speech) has long been realised. One major reason has been widely recognised as that the Lombard effect highly varies from speaker to speaker since the early 1990s. This inter-speaker variability is significant in terms of the strategy with which a speaker changes the speech intensity in noisy environments. Junqua [146] suggests that this difficulty might be reduced by identifying the subgroups of speakers in

which a common or similar adaptation strategy is shared. Some studies on the Lombard speech report the results divided by subgroups of objectives. For example, Boril [25] reports a Lombard effect equalisation scheme for speech recognition, and the results are reported for males and females respectively. On the other hand, the speaker variability makes Lombard speech perform well in a speaker identification task [108].

In speech recognition research into noise-robust techniques, recognisers need to be evaluated on noisy speech. A common way of acquiring noisy speech is to add acoustic noise to clean speech data [119] [208] [208] [48]. This approach is limited because, by injecting additive noise, it does not take the Lombard effect into account. The noise conditions used in these studies are normally measured by signal-to-noise ratio (SNR) as both the speech and noise signals are fixed before the injection. However, studies show that SNR does not have a close correlation with the recognition success [101] [234]; instead, the recognition process is more closely associated with signal amplitude. In contrast, to account for the speaker behaviour change in noisy environments, systems need to be evaluated on the speech recorded in an actual acoustically noisy environment (i.e., Lombard speech). The noise conditions used in latter studies are often measured by amplitude/sound-pressure-level (SPL), as the noise conditions are presented before the actual speech recording [26] [108] [113] [146].

Over the years, techniques have been studied to compensate for the Lombard effect in ASR. For example, an HMM-based model is used to generate Lombard speech tokens from neutral speech, and these tokens are used to re-train the ASR systems [109]. Another study [28] investigates pre-emphasis and cepstral mean normalization (CMN) and the results are compared with traditional features. A recent study [25] proposes unsupervised frequency domain and cepstral domain equalizations; the system is evaluated in digits presented in different car noise conditions. A review of the techniques can be found in the report of Hansen [108].

It should to be noted that the analysis of affective/emotional behaviours have increased attention in speech recognition researches. Similar to the Lombard effect, users' affective

states modify their speech behaviour, therefore causing mismatch between recognisers' training and practice and leading to degrading recognition. However, there are differences between Lombard speech and affective speech. For example, Lombard speech and angry speech were reported to be different, although the fact that they both have higher pitch frequency and loudness measured to neutral speech makes them difficult to differentiate [203]. A detailed review of affective recognition can be found in the report of Zeng [326]. However, the analysis of affective state is not involved in this thesis and will not be further discussed.

Besides the techniques used in speech-only systems, another approach to deal with the speech recognition degradation brought by noise environments has been to design a multimodal ASR. This is the approach taken in this thesis using gaze as an extra input modality.

2.4 Multimodal ASR

2.4.1 Types of multimodal ASR

Researchers have been trying to improve speech recognition by the introduction of extra modalities. For example, sixteen speakers were involved in a map-based interface where they were free to use speech and gesture inputs, and a 41% drop in total error rate was reported [229]. More multimodal systems that employ speech and gestural input and their modality integration methods were reviewed in a report by Oviatt [232].

Some modalities are found to be tightly related to the speech production, such as lip movements and facial gestures. A variety of studies have been conducted to exploit these modalities for improving speech recognition. These related modalities are typically sensed by cameras; therefore, the associated studies are often referred to as 'audio-visual speech recognition'. One motivation of audio-visual speech recognition is to mimic humans' nature in speech perception; humans combine audio and visual information to decide

what is being spoken. For example, hearing-impaired people perceive speech with the help of 'speech-reading' by observing mouth movement [23] [190]. A major reason visual information benefits the speech perception is that it provides complementary information of the articulations. The articulators, include tongue, teeth, lips, jaw, and lower face muscles are found to be correlated to speech production [15] [323]. The partial or full visibility of these articulators is reported to improve human speech perception [38] [298].

The visual features extracted in the first speech-reading system in 1984 [244] include mouth height, width, perimeter, and area. Inspired by the success, a variety of audio-visual speech recognition studies have reported utilising visual features, such as lip movement [40] [318], face detection [92] [91], or mouth tracking [81] [295]. However, in some studies the articulatory features are not necessarily sensed by video recording devices. For example, in a study that exploits multimodal data fusion in ASR [112], the movements of the lips, tongue, and jaw are tracked by an attached electro-magnetic articulography device. Overview of the visual features, audio-visual integration schemes and articulation knowledge in speech recognition can be found in reports by Potamianos [249] [251] and King [158].

2.4.2 Handling noise in multimodal ASR

It is not surprising that by exploiting extra visual modalities, multimodal ASR has demonstrated more favourable performances compared to audio-only systems in dealing with a variety of conditions, which include the Lombard effect [127]. Many multimodal speech recognition studies have been evaluated on noisy speech and the improvements over audio-only recognition have been reported. For example, in a study reported by Potamianos [250], mouth-region visual information is exploited and the system is evaluated on speech contaminated by various additive babble noise. In a study by Navarathna [217], drivers' lip movement is combined for in-car speech recognition, whereas in another study by Faubel [81], the in-car recognition is enhanced by mouth-localisation. In a study where the articulatory information is retrieved using an electro-magnetic articulography device,

a 64% increase in accuracy is reported in the noisiest ($-10dB$ SNR) level [112]. Interestingly, some studies have reported recognition enhancement in audio-only systems by using articulatory information extracted from noisy speech [161] [158] [202].

One important finding is that the Lombard effect is found not only to affect the speech modality, but also to have an impact on visual modalities. An effect on visual channels in 'speech-reading' is mentioned by Huang [127]. When visual speech is recorded in noisy environments, the visual features of the mouth and face also change, and the recognition rate is further improved if the visual change (visual Lombard effect) is taken into account [113]. The Lombard visual speech has been reported to have more jaw, mouth, and head movements as well as increased correlation between audio and visual modalities [58]. Where audio Lombard effect is a phenomenon with increased intelligibility to aid communication (see section 2.3.3), the results suggest that the increased intensity in visual modalities and audio-visual correlation serves the same objective. Moreover, the greater change in babble noise than in white noise in terms of more intense frequency and more jaw and mouth movements agrees with the audio-only findings where more intelligibility can be observed in babble noise compared to in white noise [146] [148].

Articulatory features are tightly related to speech due to the nature of speech production and perception. While exploiting these tightly related modalities has been showing great potential in multimodal speech recognition, the use of those modalities that are not directly involved in articulatory process has also been investigated, such as gaze [49].

2.4.3 Gaze-contingent ASR

Considering the volume of multimodal speech gaze systems, the number of reported studies in exploiting gaze for robust speech recognition in a 'gaze-contingent ASR' is relatively limited. A common approach for integrating gaze in speech recognition is to increase the probability of a word being spoken if the related object is the focus of gaze. In a study by Sarukkai [282], if a user looks at a city on a map, the city name's word score in speech recogniser will be boosted. While in a study by Zhang [327], a gaze N-best list is generated

with the rank representing the distance between an object and the gaze fixation. This gaze N-best list is used to disambiguate the similar sounding words of these objects. More recently, studies have tried to improve speech recognition by adapting language model using the gaze information. The language model perplexity improvement is reported by Cooke [50], realised using cache-based adaptation (defined later in the thesis section 7.1.5) with the cache composed by gaze events. However, minor Word Error Rate (WER) improvement is reported by Qu [257], whereas integration of gaze is shown to improve figure of merit (FOM) by Cooke [49]. The latter study also contributes by recognising spontaneous conversation between people; anticipating this natural interaction style can be employed by future multimodal systems. It is argued that the modest increase in WER can be explained by the fact that people tend to clearly speak the words associated with visual foci in the visual fields. Due to the generally good performance recognising these words, there is not much room for improvement. However, this may not be the case when the speech recognition performance is compromised (e.g., in acoustic noise). For this reason, it can be expected that the use of gaze in speech recognition is more beneficial in noisy environments. However, so far none of these systems have been evaluated in such settings.

Based on the discussion of the existing gaze and speech recognition studies above, several main concerns are proposed for more robust gaze-contingent speech recognition:

- Gaze is 'always on', i.e., not all gaze events are related to speech. Thus, it is more reasonable to use gaze events selectively with 'relevance' considered.

- The variability of the temporal relationship for integrating speech and gaze. Among the mentioned speech-gaze systems, some have reported the benefit of using co-occurring fixations/visual attentions [327], while others used the preceding ones with a variety of optimum eye/voice spans. Even in the same study where the task settings remain unchanged, inter-speaker variability is reported [155]. Thus, a well-designed integration scheme that is not only based on fixed time delay is likely to be more desirable.

- The evaluation of the system in noisy environments. For the system to be robust in natural environments, not only the additive noise contamination, but also the Lombard effect should be involved. Thus, the system needs to be evaluated on the data recorded in real-world noisy environments.

This thesis aims to address all these concerns in the context of multimodal HCI systems and gaze-contingent ASR.

2.5 Summary

In this chapter, multimodal systems in the context of HCI are described highlighting the concept of context awareness: extra information available to the system to assist in fulfilling its purpose (section 2.1).

In section 2.2, the gaze modality and its selective use in multimodal systems are described. Some commonly used gaze features, such as fixation duration, saccade length, pupilary response, and visual attention are discussed. The top-down and bottom-up analytical approaches and the concept of gaze roles with the underlying aspects of context awareness is presented with the discussion of the measurable validity; roles related to cognition lack external validity whereas those related to interaction and environment change can be validated. The use of machine learning techniques in gaze data analysis is discussed.

In section 2.3, the broad historical progression of the ASR is discussed with the techniques applied in this study. The impact of the acoustic noise is discussed in the context of ASR, multimodal ASR, and gaze-contingent ASR (section 2.4) respectively. The speech behaviour change in acoustic noise (termed as acoustic Lombard effect) is discussed with the change in other modalities.

Gaze is an underutilised modality in the ASR systems, and the related studies are presented with the expectation that the use of gaze is more beneficial in acoustic noise.

CHAPTER 3

INTEGRATION OF GAZE AND SPEECH IN ASR

In section 2.3 and 2.4.3, it has been discussed that the performance of ASR in acoustic noise can be improved by:

1. The knowledge of the noise condition.
2. The information in gaze information events (e.g., visual attention) that are related to spoken words.

In this thesis, these two concerns are addressed with a proposed architecture described in section 3.1. A working gaze role taxonomy is proposed in section 3.2. The coupling of gaze events with other modalities is described in section 3.3 with the implementations in acoustic noise and visual attention inference outlined in section 3.4 and 3.5.

3.1 Architecture

Figure 3.1 shows the design architecture of a selective gaze-contingent ASR system. As introduced in section 1.6, the selective use of gaze in the ASR is achieved by acoustic model adaptation based on acoustic noise inference (ANI) and the language model adaptation based on visual attention inference (VAI).

The ANI will be realised by quantifiable measurement of the relationship between speech and gaze and the dependency of the relationship on acoustic noise. The ANI

framework is outlined in section 3.4 and evaluated in Chapter 5.

In section 2.2.3, it has been discussed that visual attention is a specific type of gaze information event, which is closely related to the psycholinguistic processes. Visual attentions have been exploited in ASR systems to improve recognition performance [49] [50] [257] (see section 2.4.3). However, as gaze is 'always on', it is reasonable to assume that some visual attentions are more related to spoken words (i.e., relevant to system function in ASR systems) than others. In this study, it is proposed that the use of visual attentions be selective: those related to speech should be used in integration rather than all visual attentions. The visual attention inference (VAI) will be achieved using gaze events characteristics and the coupling with speech and system response events. The VAI framework is outlined in section 3.5 and evaluated in Chapter 6.

The speech used in both inference frameworks is a baseline ASR output without introducing gaze input. A selective gaze-contingent ASR that utilises both inference frameworks is built and evaluated in Chapter 7.

3.2 Taxonomy of Gaze Roles for ANI and VAI

In section 2.1.4 and 2.2.4, the cognitive, interactional, and environment reaction gaze roles and the underlying aspects of context awareness (CA) are discussed. These complementary descriptors for gaze behaviours - cognition, interaction, and reaction to environment - afford simultaneous interpretation, i.e., behaviours relating to cognition do not cease when behaviours with an interaction are inferred and vice-versa.

Building on the previous work of various gaze types [266] [292] [221], a working (non-exhaustive) taxonomy for gaze roles in this thesis is proposed:

- *Cognition Roles.* Gaze behaviours relating to cognition that lack validity by direct measurement, although evidence for their existence in cognitive science is strong (see section 2.2.4). Examples include object naming, mediating attention, reading, scene perception, visual memory, sentence planning, and other psycholinguistic roles.

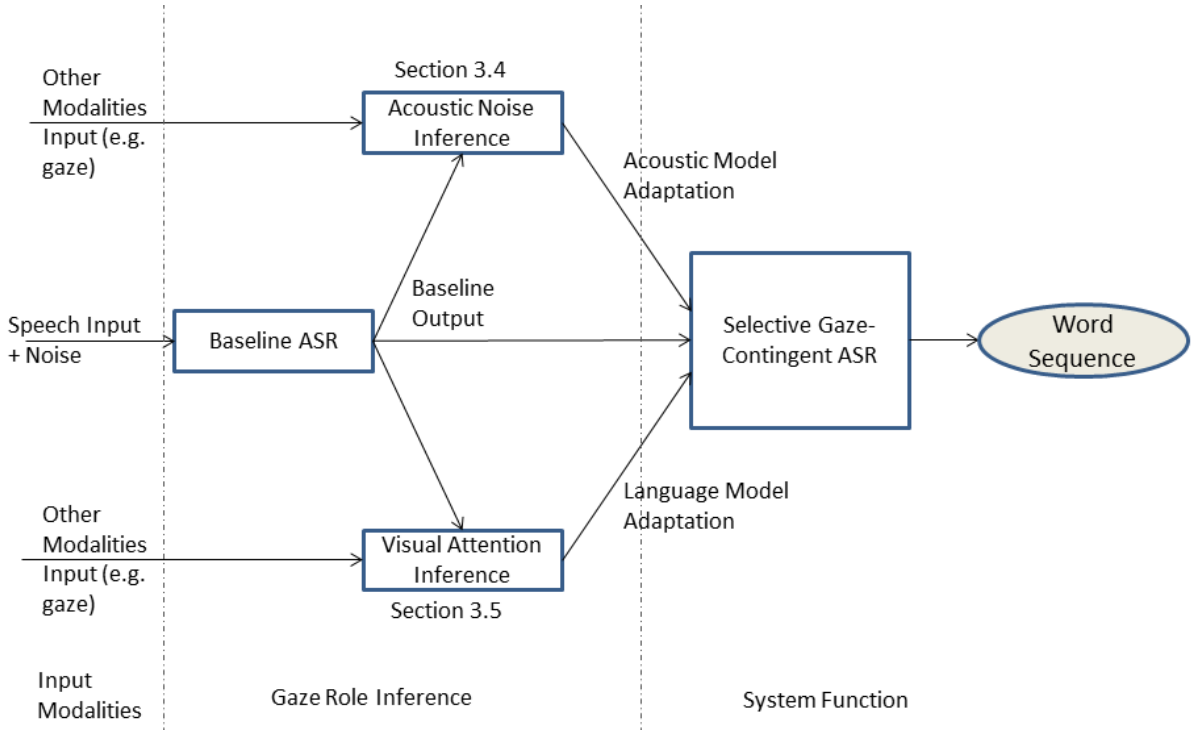


Figure 3.1: The system design architecture for inference of acoustic noise condition and relevant visual attentions. The selective gaze-contingent ASR benefits from utilising the inference results. ANI results are used for acoustic model adaptation and VAI results for language model adaptation. The arrows represent the data flow.

These cannot be inferred directly but, as is shown in Chapter 5, can be utilised to infer the acoustic noise condition in ASR.

- *Interaction Roles and Environment Reaction Roles.* Gaze behaviours relating to interaction and reaction to changes in the environment; directly measurable. These can be inferred directly following the probabilistic Bayesian approach outlined in Chapter 6 and include:

  - *Task-oriented Visual Attention*(TOVA) - Gaze behaviours associated with tasks and activities assumed by the system. Examples include typical gaze-oriented roles, such as eye typing, selection, and gesturing.

  - *Social Visual Attention*(SVA). Gaze behaviours associated with social interaction between people. Examples are establishing agency, regulating interaction, communicating social attention, and conveying emotion.

  - *Reactive Visual Attention* (RVA). A person's gaze guided by changes in the environment - e.g., visual or auditory disturbances, such as changes in displays, noises and so on.

  - *Task-independent Visual Attention* (TIVA) - Gaze behaviours other than the above ones that do not contribute to the system task or interaction.

The taxonomy splits context awareness (see section 2.1.4) into three categories - cognitive context awareness (CCA), interactional context awareness (ICA), and environmental context awareness (ECA) (Figure 3.2). Each CA category contains a number of hypothesised aspects. For CCA, these are related to the hypothesised cognitive process, i.e., 'what the user is thinking'. For ICA, the aspects relate to how people interact with one another, i.e., 'what the user is communicating'. For ECA, this relates to 'traditional' CA aspects needed to infer gaze behaviour, such as the state of the visual field and acoustic noise. The aspects in each CA category are not exhaustive and, depending on system function (and current 'best practice'), will differ. It needs to be noted that, while ICA

and ECA can be observed and measured, CCA (e.g., brain activities) cannot be observed and are thus considered latent/hidden/confounding variables.

The two cognitive gaze roles exploited in this study is object naming - where people look at objects prior to naming them, and mediating attention - where people look at objects during naming them. These preceding or co-occurring gaze events are commonly utilised in gaze-based multimodal systems to integrate with speech (discussed in section 2.2.6 and 2.4.3).

Task in TOVA refers to a task that the user is undertaking, which is *assumed by the HCI system*; for example, it may be a hypothesised task relating to looking at the spoken objects to assist speech in noisy environment (e.g., in a 'put-that-there' task where a gaze-contingent ASR can be exploited). It is reasonable to assume that different tasks will require different use of CA and thus manifest different gaze behaviours; i.e., depending on system function, there may be more than one TOVA. The visual attentions unrelated or not contributing towards the system function are described by TIVA.

Social visual attention (SVA) describes gaze behaviours associated with social interaction. Examples include establishing agency, regulating interaction, communicating social attention, and conveying emotion. These interactive contexts are taken from work in eye movement synthesis in robotics [292]. In terms of an HCI system where only one user is involved and no social interaction is assumed, SVA is not considered in the scenario.

Reactive visual attention (RVA) occurs where a person's gaze is guided by changes in the environment - e.g., visual, auditory or oflactory. An example of RVA would be a visual attention reacting to the change in the display.

3.3 Coupling between Gaze And Speech Events

Inferring roles for gaze by measuring gaze features only may not be the best practice because many roles (shown in Figure 3.2) are associated with multimodal behaviour. Therefore, inference of gaze roles requires defining the information events and features

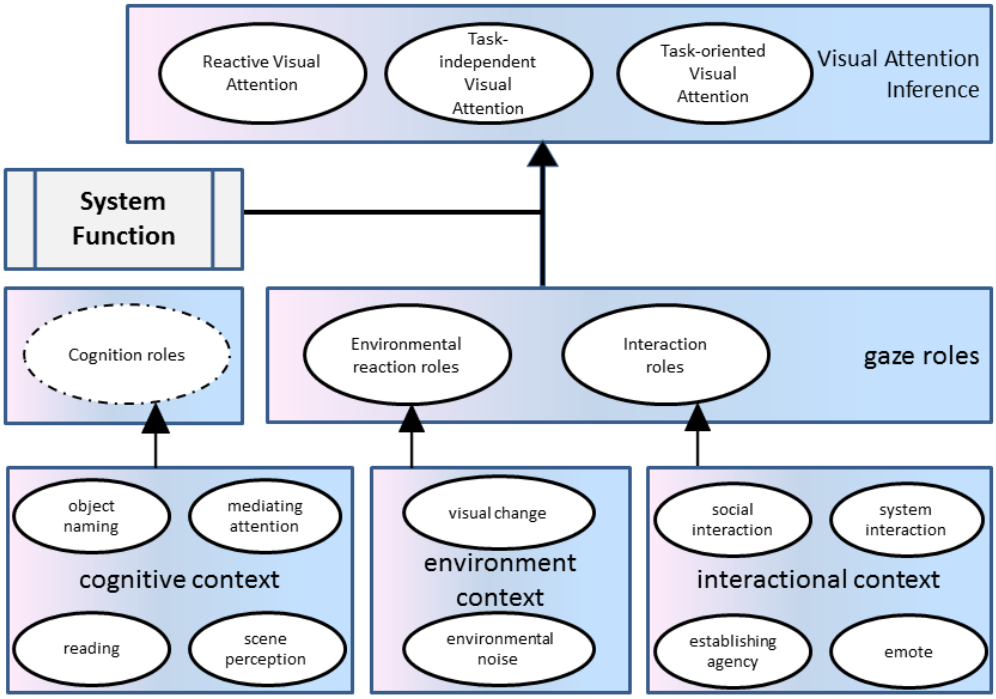


Figure 3.2: Proposed working gaze role taxonomy for this thesis. The arrows represent that three classes of gaze behaviour are defined related with the underlying CA from cognitive, environment reaction and interactional processes and the visual attentions inferred in this work is related with the measurable gaze roles. Elements in different boxes are treated differently within the working taxonomy. Task refers to the user's task assumed by the HCI system.

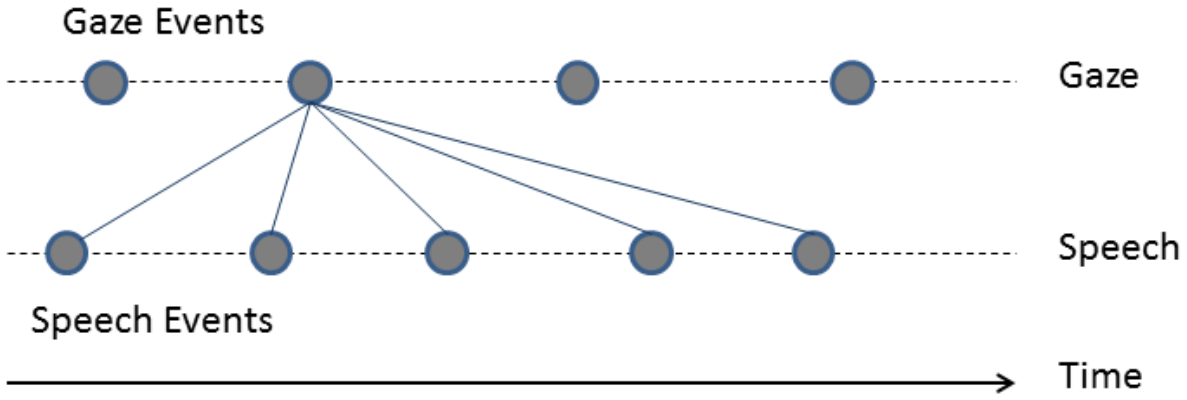


Figure 3.3: Proposed coupling framework between information events in different modalities.

to be recognised in gaze and other modalities: e.g., for gaze, the events are information occurrences that are detected by measuring gaze features, such as fixations and their durations upon identified objects, and for speech, events could be words. The events in other non-verbal human modalities could be defined by established taxonomies, e.g., facial expressions signifying specific emotions, body gestures signifying deictic reference and so on. Modalities could also be other external sources of information, e.g., changes to a display or the actions of another person. Of the interest in this work, the coupling approaches are demonstrated specifically between gaze and speech modalities.

As part of the inference, the semantic and temporal relationship is defined between these events. The semantic relatedness defines how strong events across modalities are coupled in terms of their role-dependent meaning. The temporal relationship between the events defines how strong events across modalities are related according to their occurrence in time. In section 2.4.3, the variability of the temporal relationship between speech and gaze for integration is discussed. A coupling framework is proposed with all events in one modality coupled to all events in another within a temporal window. Figure 3.3 shows the coupling between a gaze event and all those in speech. Each couple has a 'weight' or 'strength' determined from temporal and semantic constraints.

Thus, the role(r)-dependent multimodal coupling between two modalities $f_r(.)$ can be

defined to be the semantic relatedness $f_r^s(.)$ and temporal relatedness $f_r^t(.)$ combined using a certain function $h(.)$. More formally, the semantic and temporal relationship between information events in gaze and speech can be stated using a general coupling function:

$$f_r(G, W) = h(f_r^s(G, W), f_r^t(G, W)) \tag{3.1}$$

where $f_r^s(.)$ and $f_r^t(.)$ are functions that measure the semantic s and temporal t relationship respectively between gaze event sequence $G = (g_1, g_2, \ldots, g_t)$ and speech event sequences $W = (w_1, w_2, \ldots, w_v)$ based on a defined role r. The function $h(.)$ is the means by which temporal and semantic functions are combined. For the two types of gaze role defined (cognition and interaction) in section 2.2.4, how this coupling function is realised will differ as described and implemented in Chapter 5 and 6. It needs to be noted that in event-based frameworks where only single events in the sequences are concerned, for example gaze event g_t and speech event w_v, the G and W in the expression 3.1 can be replaced by g_t and w_v respectively.

3.4 Mutual Information And Its Utilisation in Acoustic Noise Inference (ANI)

As discussed in section 2.2.4, gaze roles relating to cognition cannot be measured, and thus, cannot be directly inferred. However, their relative prevalence over the temporal window can be estimated with measures for each role based on a-priori assumptions of information events and their coupling. These measures can be combined to form a feature vector used by a classifier to infer measurable variables such as environmental conditions like acoustic noise level.

Since Peng [243] investigated the use of Shannon's mutual information (MI) [285] in calculating correlation between features for the feature selection, countless studies adopted

this MI approach to study the relationship between features. For example, being non-linear, MI is reported to be effective in representing the relevance and dependencies of features in data mining [184] and pattern recognition [80]. MI is also used to combine multi-sensor data inputs within a statistical framework, and the main advantage is shown to be the capability of measuring various relationships between information sources taking into account the underlying uncertainty [201]. However, none of them investigated the value of MI measure in studying the relationship between gaze and speech to improve ASR robustness in acoustic noise.

In the previous studies, the MI measure has been compared with other correlation measures. Roy [273] [275] built word acquisition systems based on maximum MI between co-occurring spoken and visual input to combine similarity metrics without the need for ad hoc heuristic measures. Estévez [80] claims that the MI measures distinguish it from other correlation measures in two main properties: first is the capacity of measuring any kind of relationship between variables and second is its invariance under space transformation [171]. MI is also used to measure the non-verbal human-robot interaction for the behaviour recognition system and is compared with conditional entropy(CE) and Kullback-Leibler divergence(KL) [196]. For two random variables X and Y, the CE represents the uncertainty of Y given X, averaged over all possible values for X, and KL represents the similarity between X and Y distributions. His results state that, although MI measures the relevance/dependency of X and Y and is closely related to CE and KL, MI is a significantly better measure for assisting interaction prediction.

To quantify the coupling of gaze events with speech over a temporal window, the measures proposed utilise MI. Gaze and its coupled modalities are modelled as a sequence of discrete-valued random variables representing the events that define their cognitive role, such as object naming and mediating attention (see section 3.2). Let g and w be the random variables for gaze event (e.g., visual attention) and speech event (e.g., word sequence) respectively. The MI $I(G;W)$ is a measure of the difference in entropy between the joint density $p(g,w)$ and the product of the marginal densities $p(g)$ and $p(w)$:

$$I(G;W) = \sum_{g \in G, w \in W} p(G=g, W=w) \log_2 \frac{p(G=g, W=w)}{p(G=g)p(W=w)} \tag{3.2}$$

An MI measure can be proposed for a specific relationship between gaze and speech. An MI of $0bits$ (assuming base 2 log) indicates that the modalities are uncoupled - i.e., that the event occurrence in one modality is independent of the event occurrence in another modality. An MI $> 0bits$ indicates the strength of the assumed relationship. For a random variable pair $(X;Y)$ of n possible values with possibilities $p_1, p_2, \ldots, p_n$, the theoretical maximum value of the mutual information $I_m(X;Y) = log_2^n$ when $p_1 = p_2 = \ldots = p_n$ (e.g., $I_m \approx 3.20$ when $n = 9$). The multimodal coupling function (expression 3.1) is used to estimate the joint and marginal densities from the coupled event occurrences in the gaze and another modality over a window of time T. That is saying, the densities are estimated based on how multimodal coupling functions are defined:

$$f_r(\{g_{t-T}, \ldots, g_t\}, \{w_{t-T}, \ldots, w_t\}) \tag{3.3}$$

MI can be interpreted as a generalised measure of relationship strength for it being analogous to other correlation measures, but sensitive to the functional relationship defined by the coupling function, not just linear dependencies (e.g., Pearson correlation or Euclidean distance) [294]. Of interest for this study is, in contrast to those more commonly used measures that quantify linear dependencies [294], MI is 0 if and only if the gaze and speech for the specified relationship is statistically non-related/independent. The implementation of the ANI framework and the analysis of the MI approach results will be discussed in Chapter 5.

3.5 Visual Attention Inference (VAI)

Interaction gaze roles can be directly inferred; sequences of multimodal data are labelled with the gaze role according to rules, taxonomies, and behaviour patterns. As for noise inference, the multimodal coupling function in expression 3.1 may be defined as corresponding to a relationship between speech and gaze. The coupling function can also be defined to assist the gaze role inference. The difference is that the function is used in inference to define a likelihood function rather than a feature vector.

Applying Bayes theorem for classification, the probability that a gaze event can be attributed to a gaze role r at time t is proportional to the product of the likelihood function and prior:

$$P(r_t = r|g_t = g, W) \propto P(g_t, W|r_t = r)P(r_t = r) \tag{3.4}$$

which can be rewritten as:

$$P(r_t = r|g_t = g, W) \propto P(g_t = g|r_t = r)P(W|r_t = r, g_t = g)P(r_t = r) \tag{3.5}$$

The liklihood $P(g_t = g|r_t = r)$ is described by probability density function (pdf) $p(g_t|r_t)$ representing gaze event characteristics. As a demonstrative approach, the likelihood pdf $p(W|r_t, g_t)$ can be calculated by estimating the relationship between a gaze event g_t and speech sequence W for role r via the coupling function $f_r(g_t, W)$ normalised with respect to the coupling functions for other roles so the total probability $\sum_{v=1}^{V} P(W|r_t = v, g_t = g) = 1$:

$$P(W|r_t = r, g_t = g) \propto \frac{f_r(g, W)}{\sum_{v=1}^{V} f_v(g, W)} \tag{3.6}$$

where $v \in V$ denotes the set of considered roles.

Note that for systems inferring a gaze role from gaze behaviour alone, expression 3.5 reduces to:

$$P(r_t = r|g_t = g) \propto P(g_t = g|r_t = r)P(r_t = r) \tag{3.7}$$

The implementation of the VAI framework will be discussed and evaluated in Chapter 6.

3.6 Generalising Speech to Other Modalities

In this chapter, the coupling functions are described between gaze G and speech W specifically. As discussed in section 3.3, when another modality M is considered (e.g., system responses) to be related to gaze role behaviours, the speech sequence W and event w in the proposed frameworks can be replaced by the sequence of this modality M and event m respectively.

3.7 Summary

In this chapter, a fundamental coupling framework of the information events in two modalities is described. The application in the ANI framework for acoustic adaptation is presented with the MI being a relationship measuring approach based on cognitive gaze roles. A general framework of VAI in the integration process for language model adaptation is discussed based on the gaze roles associated with interaction and reaction to environment change. The implementations of the frameworks in this study will be described and evaluated respectively in Chapter 5 and 6.

CHAPTER 4

A CORPUS OF GAZE AND SPEECH IN ACOUSTIC NOISE

From about 3 or 4 years of age, we have the ability to express and comprehend the speech and gaze behaviours of others [65] [207]. In Chapter 2, it has been discussed that in an acoustically noisy environment, we adjust our speech and gaze behaviours for communication with others [204]. Interactive systems increasingly employ natural multimodal interaction styles and operate in real-world environments, where acoustic noise could be involved. For such systems to interpret a user's speech and gaze to understand the communication intent behind the modalities (i.e., instructions the user wants to express to the system), the behaviours of the user's gaze and speech, their relationship, and how those change in noise need to be learnt. For this purpose, a corpus of speech and eye-tracking data is collected in different acoustic noise environments.

This chapter describes the experiment set up and the data collected, which will be referred as the 'Eye-Speech-in-Noise (ES-N) corpus' hereafter. The concept of the Wizard-of-Oz (WoZ) simulation and the motivations of using the setting are discussed followed by its implementation. The experiment task implementing WoZ simulation for the speech-gaze data collection is described. The participants and apparatus of the experiment are introduced.

4.1 Motivation for Collecting the ES-N Corpus

Corpora are collected for speech technology and linguistic researches. These corpora are collected to benchmark systems and provide standard for research evaluation and so on. In an early study Doddington [64] pointed out that to turn speech recognition theory to practice, formal evaluation tests need to be performed on real data. Consequently in his study, a speech corpus with a vocabulary of 10 digits and 10 command words was collected for the evaluation. The TIMIT corpus [96] of read speech was collected to provide data for the development and evaluation of automatic speech recognition systems. TIMIT contains speech representing 8 major dialect divisions of American English with 10 phonetically-rich sentences spoken by each speaker. The Wall Street Journal-based continuous speech recognition (WSJ-CSR) corpus [241] was used in the benchmark tests for the APRA spoken language program [238].

Some corpora are read-speech corpus, such as broadcast news [227], lists of words, or sequences of numbers [176] and also there are spontaneous-speech corpus, such as dialogs (between two or more interlocutors) [289], map-tasks (one explains to another) [5], appointment-tasks (two persons try to reach a common meeting schedule) [140], or narratives (story-telling) [247]. Each corpus may contain English or other foreign languages, such as Chinese mandarin [303]. The amount of multimodal speech corpora is increasing to support the high-level understanding in the context of meetings. For example, VACE [41] captures multimodal data, such as speech, gaze, gestures, and postures, for understanding meetings, while AMI [194] is a corpus containing data of speech recognition, computer vision, dialogue, discourse, and meeting abstraction for developing remote-meeting assistants and meeting browsing frameworks.

There are existing speech and gaze corpus in the human-computer-interaction systems. For example, the Cooke's eye/speech corpus [49] contains speech and synchronised eye movement data collected through a collaborative map describing tasks involving two participants. In total, 9 participants were involved, and 18 sessions were recorded using 9 maps with each session lasting for 5-15 minutes. The collection of the Cooke's corpus was

motivated by the evaluation of the gaze-contingent automatic speech recognition (ASR) system built for the study. However, the amount of such corpus is very limited, not available publically, and does not specifically pay attention to the effect of acoustic noise. And as far as the author is aware, there is no such speech-gaze corpus that is recorded in different levels of acoustic noise.

The motivation of collecting the ES-N corpus is to gain practical data to explore the noise-dependent relationship between speech and gaze. The corpus has value for the evaluation of future speech-gaze researches.

4.2 Wizard-of-Oz Simulation

An experiment task is set up using the Wizard-of-Oz (WoZ) simulation paradigm described in this section.

4.2.1 WoZ as a research method

Wizard-of-Oz [55] or high-fidelity simulation [193] is a methodology used in intelligent systems for the simulation of high-level functions. A human so-called 'wizard' simulates a system that interacts with the human users just like the envisioned system. Ideally, the users would not be aware of the simulation; therefore, they behave as if they were interacting with a real system. A general objective of using WoZ is to simulate the system components or functions that require the most effort or are not feasible with the on-going technology (e.g., accurate speech recognition in noisy environment).

The WoZ simulation has been used in various applications since the 1970s. For example, in an early study [193], an intelligent agent is simulated to learn actions, pointing at or talking about relevant data from the users. Later examples include the experiments for collecting speech and gaze related data in interactive systems. In a recent study [129], a WoZ system is presented for predicting human attention interruptibility using the audio and video data sensed, and a 78% accuracy is reported. Another study [222] is reported

where a gaze-aware 'look-to-talk' interface is developed for the users to interact with an animated virtual agent, and the WoZ system is compared with a real recognition system in terms of user preferences. WoZ paradigm is also used for collecting the speech and gaze data for a dialogue system [296].

In a study that applies WoZ simulation in service robots [102], the data obtained from the WoZ study is stated to be qualitative to a large extent in terms of comparable user experience to regular interfaces. An important feature of the WoZ simulation is allowing the system designer to act as the system (i.e., become a wizard) because during the interaction, the wizard is personally responsible for the user's experiences and confusions, which motivates and justifies the revision [193]. Sometimes this is also the cause of particular concerns, such as whether the data can represent the real system utility so that the validity of the simulation setting needs to be considered based on the purpose of the system [193].

4.2.2 Motivation

A challenge with designing an interactive system that employs a natural multimodal interaction style is to capture a user's behaviour when he/she interacts with it. Because such intelligent systems are expensive or performances are not yet adequate, there is a need to simulate the system using WoZ settings.

With the WoZ setting, the designer can benefit from the rapid set up of the system to conceptualise and attempt new interface ideas. In addition, important early design experience can be gathered through the preliminary implementation phase [103], and an impression of the system interactions may be learnt from early user experience.

By acting as the 'wizard' and interacting with the users, the designer can improve the understanding of what kind of interaction the system should support or the suitability of the modalities. The typical results produced by the WoZ studies can be the corpora in various formats for different purposes (e.g., linguistic data for language modelling and speech recognition [6]).

Of the interest in HCI studies, WoZ supports the premise that computers can be considered as a social actor with similar interaction methods to people. While this is generally believed to be an important advantage of the WoZ setting, it has also been criticised for failing to elicit the same user behaviours observed with practical systems (limitations of real system) [102]. Thus, the purpose of the system - particularly the premise that it is a natural social actor - needs to be considered as well as the allocation of the simulated components in the system.

The validity of using WoZ simulation can be assessed by comparing the humans' task performance with machines'. Fitts [86] reported a list asserting what 'men are better at' and what 'machines are better at'. The list was further discussed and developed by Sheridan [288]. According to their studies, machines are better than humans at responding quickly to signals, reasoning deductively, and storing and erasing information. Humans are better at detecting small changes in the environment, perceiving patterns, improvising, and making judgments. A WoZ system that recognised gaze and speech was compared to a real system, and the worse performance of the real system resulted in less natural interaction with users [222].

Considering the objectives of the corpus collecting task in this study (see section 4.1) are to collect practical data for exploring noise-dependent speech-gaze relationships and speech-gaze-based system evaluation, there are several reasons to adopt the WoZ simulation. First, the interactive system is hypothesised to employ natural multimodal interaction in a real-world utility (i.e., to be human-like). Secondly, as the effect of acoustic noise is a key part of the thesis, a wizard can better respond to its change in the environment. Most importantly, a wizard possesses the capability to better perceive a user's speech and gaze behaviour in the acoustically noisy environment, while no real system can achieve this yet. In addition, the designing of such a system cannot be realised without the study of the corresponding data.

4.2.3 Abstract system descriptions

Abstract descriptions of the system are used as a guide for the wizard to prepare the interaction and for the users to realise the extent to which they can naturally interact with the system. This information sets constraints not only on the condition of use, but also on the focus of the system [102].

The three kinds of abstract descriptions are considered prior to the implementation:

- **User instructions** provide the information to the users about how they are supposed to interact with the system and/or the tasks they must perform.
- **User behaviours** hypothesise the behaviours that the designer believes the users will perform during the task.
- **System behaviours** specify the function of the wizard in the interaction, how the wizard is supposed to control the system, and the feedback to give.

This information is supposed to aid both the wizard and the users. A wizard should be trained to answer these questions before the real data collection experiments. In a large-scale multimodal interaction context, it is typical to use two or more wizards to reduce the amount of work each wizard undertakes, such as the dual-wizard scheme used in a framework for designing a speech-pen system [46]. However, to maintain the appearance of a single system, all wizards need to be trained properly to have a similar understanding of the system and collaborate closely to act as one system. To remove the potential bias caused by different understandings of these questions, in this study, the same wizard is used in all data collection tasks. The implementation of these abstract descriptions for the experiment will be described in section 4.3.2.

4.2.4 Data collection and pilot study

One of the main purposes of conducting a WoZ simulation in the context of HCI research is to collect data of user behaviours. With a proper setting that meets the designer's

requirements, a large amount of data can be collected. However, this is expensive, and it is important to decide upfront what type of data to collect and the parameters or variables setting to meet the purpose of the study. As mentioned previously in section 4.2.2, one advantage of WoZ is allowing quick system set up; a pilot study for initial data collection can be conducted to fulfil these objectives.

4.3 Method

In this section, the experiment task implementing WoZ simulation for the ES-N data collection is described. The participants and apparatus of the experiment are introduced.

4.3.1 Task

The task is a collaborative puzzle task based on Schmandt's seminal 'put-that-there' [283]. The task involves a user utilising gaze and speech to instruct the system to position a coloured shape on a map displayed on a computer screen. This means that during a task, the user needs to convey three elements to the system - colour, shape, and position - to form a 'spatial location' instruction. The envisioned system perceives the user's speech and gaze direction and gives responses by updating the result on the displayed map based on the perceived instructions. The task ends by the user confirming the correct outcome on the screen. In this task setting, the user's instructions to position a certain coloured shape is so-called his 'communicative intent'.

Users do not always interact multimodally in a multimodal system. In this task, they use speech and gaze to issue spatial location instructions. Gaze is optional, and the instruction could be issued by speech only. However, in a previous work [228], it was revealed that 95% to 100% of users preferred to interact multimodally when they were free to use either modality in a spatial domain, while the figure was below 70% in numerical domains and 60% in verbal domains. In this task, the map is therefore designed for maximum spatial separation between elements (Figure 4.1), and the interaction objectives

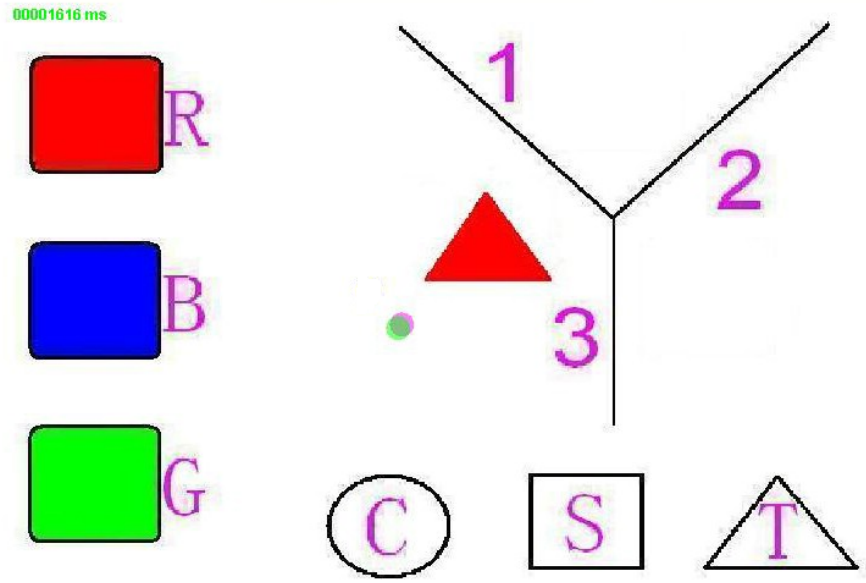


Figure 4.1: The spatial puzzle task used to elicit different gaze behaviours in acoustic noise. The user instructs the system to place a coloured shape in one of three locations (in this case a red triangle in location 3/left). The user's focus of visual attention is overlaid on the task map (small circle).

are clearly laid out across the map to encourage the participant actively using both speech and gaze to convey the instructions.

The task is repeated for different combinations of coloured shapes and positions selected at random. The task is undertaken under different acoustic noise conditions, including a no-noise (baseline) condition. More details regarding the noise conditions are discussed in section 5.3 and 4.5.1.

4.3.2 WoZ implementation and system descriptions

In the WoZ simulation setting, one or more 'wizards' act as one envisioned system (see section 4.2.3). In this work, a single wizard is used to remove the potential bias caused by different understanding of the task and personal operating style.

The wizard on behalf of the system responds to the user's instruction by updating the result on the screen or vocal (optional) feedback. The wizard terminates the task upon receiving the user's confirmation of the correct result. An example of a task process

```
User: Place the blue circle on the top.
Wizard: Green circle on the top?
       (Place green circle on top)
User: No, blue. Blue one.
Wizard: Alright, blue circle on the top?
       (Place blue circle on top)
Wizard: Like this?
User: Yes, good job.
Wizard: OK.
       (End the task)
```

Figure 4.2: A sample dialogue for the spatial puzzle task. The user tells the wizard to place a coloured shape in one of three locations (in this case a blue circle in top location).

is listed in Figure 4.2 assuming the wizard provides vocal feedback (the actions in the brackets are the updates on the screen controlled by the wizard):

With the knowledge of the hypothesised system, the three abstract system descriptions (see section 4.2.3) are:

- **User instructions:** *'Tell the system the colour and the shape you choose and the position to place this coloured shape. The system will try to understand your instruction from listening to your speech and watching your gaze.'*

- **User behaviours:** The user will issue the instruction in front of the computer screen. The user will use speech to issue the instructions that involve choosing a colour, a shape, and a position. With the knowledge that the system is monitoring gaze, the user may use it to aid speech. The user's use of speech and gaze depend upon the acoustic noise in the environment.

- **System behaviours:** The system updates the results on the screen according to the perceived instructions. The system optionally gives vocal feedback. The system terminates the task when the user confirms that the result is correct.

These descriptions aid the understanding of the data collection system before the actual experiment is conducted. They point out to the users the modalities they ought to be using during the task and serve as a guide to prepare the wizard.

4.3.3 Experimental procedure for corpus collection

Two experiments are conducted for the corpus collection; a pilot study for initial data collection is performed prior to the main data collection experiment.

The pilot study addresses several issues. First, it tests the apparatus to use for the data collection and provides a general impression of the experiment processes. Secondly, it is used to determine which acoustic noise conditions (motivated in section 4.1) are best used to elicit the change in gaze and speech behaviours. It aims to explore the quality of speech and gaze data collected with the equipment and some initial analysis on modelling the relationship between gaze and speech. Also, the practical experiences from the users can serve as an informal usability test of the WoZ simulated system. The findings of the initial data collection are described in section 5.3, and the system used in the main data collection is adjusted accordingly.

4.3.4 Participants

The task is repeated multiple times. Each repetition is referred to as a 'session'. Each session involves two participants: a user issuing spatial instructions and a wizard acting as the system to react to the instructions. The user's instruction mainly involves a combination pattern that defines 3 elements: colour, shape, and position described in section 4.3.1. The user is the observation target and the source of data recording.

Four participants took part in the initial exploratory experiment, and another seven participants took part in the main corpus collection experiment as the users. All participants are PhD students from the University of Birmingham. There are 8 males and 3 females with ages ranging between 23 to 26 and Chinese as first language. Their spoken English is sufficient enough to fulfil the task requirement. Participants are randomly picked from the volunteered candidates. All participants do not have previous experiences with the gaze-tracking experiments nor the equipment used. A trained wizard undertakes the role through all sessions to maintain the consistency in the system operation style.

4.3.5 Apparatus

Participants talk to one another through microphones and closed-back headphones. The headphones play both voices to each person; they hear their own voice as they speak because this has been demonstrated to be a key component in how people regulate their speech due to background noise [173]. To record the data in noisy environment, acoustic noise is added to the speech heard by participants through the headphones. Users' 'clean' speech is recorded on separate audio channels at the CD sample rate quality of $44.1kHz$ to facilitate ease of transcription. The user and wizard's microphones are studio quality and desk mounted (model 'Shure SM48'). To ensure the sound level heard by the wizard is identical to that heard by the user, an audio splitter - 2x Jack $3.5mm$ stereo (female), 1x Jack $3.5mm$ (male) stereo - and two identical headphones are used. The fact that the wizard can also hear the noise makes it possible for him to misrecognise the user's instruction.

The visual task displayed on the computer screen is augmented at time t with the user's focus of visual attention at time t so that the wizard can see what the user is looking at. The user's map is not augmented with his/her own focus of attention to avoid the phenomena of the user 'cursor chasing' his/her own gaze, the bane of many eye-typing applications [189].

The user's focus of visual attention is captured using a head-mounted eye-tracker (SR Research Eyelink 2) that captures binocular eye position at $500Hz$ using dark pupil and corneal reflection methods from infra-red illuminated video of the eyes [67] and the corresponding fixation/saccade events. More Eyelink 2 details are given in table 4.1. The accuracy enables reliable detection of fixation and saccades to positions of 2° visual angle.

The set up of the eye-tracker system involves two PCs, an Eyelink host PC and a display PC (Figure 4.3). The host PC performs real-time eye-tracking, gaze position computing, and speech recording in addition to eye-tracker configurations and performance monitoring. Online detected gaze events, such as fixations and saccades, are stored on the host PC and sent to the display PC through the Ethernet. The display PC provides the

Binocular Sampling Rate	$500Hz$
Average Accuracy	$0.5°$
Saccade Event Resolution	$0.05°$ microsaccades
Pupil Size Resolution	0.1% of diameter
Blink Recovery Time	1 msec
Gaze Tracking Range	$40°$ horizontally, $30°$ vertically
Allowable Head Movement	$+/-30°$ display
Optimum Camera-Eye Distance	$40-80$ mm
Headband Weight	420 grams
Headbank Cable Length	4.2 meters

Table 4.1: SR Research Eyelink 2 eye-tracker technical specifications

applications to design the experiment flow, view the recorded data, display calibration targets, and on-going experiment. The displayed gaze events and positions are received from the host PC.

The wizard controls the shifting of different colours, shapes, and positions by pressing the corresponding keys on the keyboard connected to the host PC. The key to choose each element is clearly stated on the task map, and the system descriptors are listed so that the wizard does not need to memorise them. The wizard changes the combination until the user is understood. When the users are satisfied with the result on the screen, they press space on the keyboard to end the task.

4.3.6 Experiment design application

The tasks are built using the SR Research Experiment Builder v1.4.402. It is the software used to create and run the task and control the recording of gaze and speech.

Within the software, a hierarchical organisation of events is defined to describe the experiment flow (Figure 4.4). The experiment is broken down to several levels along the hierarchy. This allows the task to be repeated. In the experiment, the Experiment level contains an instruction screen followed by a sub-level named as the Block level. Each Block level is undertaken under a specific noise condition several times and repeated with other noise conditions. Within each repetition of the Block level, the user performs a camera adjustment, calibration, and validation, and then runs several trials (sessions).

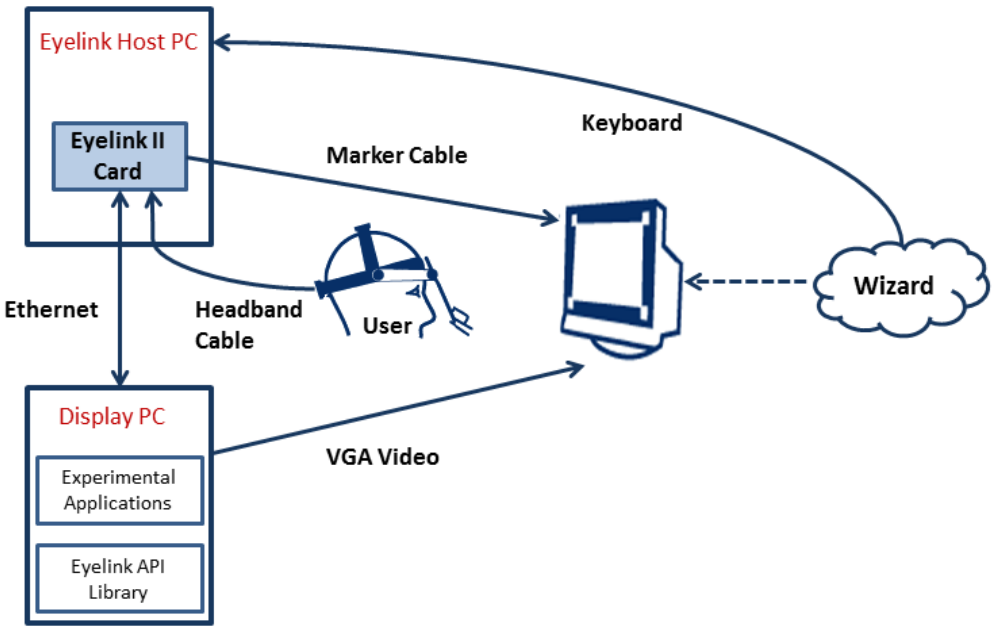


Figure 4.3: The set up of the eye-tracker system. Online detection of the gaze events and position is performed by the host PC, and the data is sent to the display PC through the Ethernet. The wizard acts as the envisioned system and reacts to the user's instructions by updating the displayed result using the keyboard.

Every iteration of the Trial level starts with pre-recording preparations (e.g., clearing trigger data, flushing log file, preloading video and audio resources) and drift correction followed by the trial recording. The Recording level is responsible for collecting the gaze data and is where the task map is presented.

Figure 4.5 shows an example layout of the software user interface. As a general step, eye-tracker initialisation is defined in the beginning of the task. After the recording starts, the coloured shape will be placed in different positions by wizard pressing 1, 2, or 3. The colour will be shifted by pressing b, r, or g, and the shape by pressing t, s, or c.

4.4 Post Processing

Post processing is performed after collecting the raw data. The gaze sequence is recorded by the eye-tracker, and a speech clip is collected through the microphone for each task session. The gaze sequence is recorded as fixation events and saccade events in-between

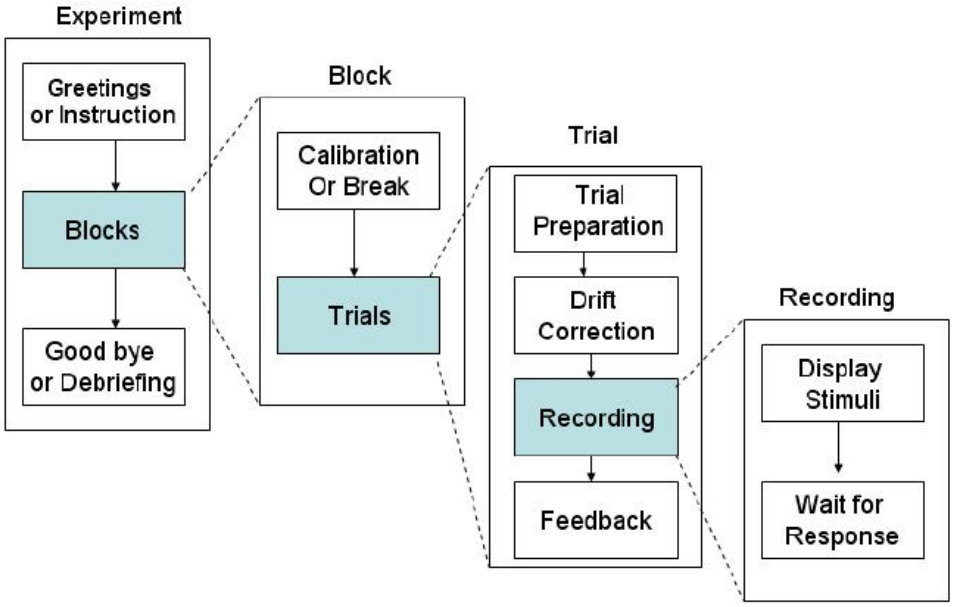


Figure 4.4: Hierarchical organisation of events used in the experiment design flow. Functions of the experiment are represented by the boxes. Arrows represent sequence.

(see section 2.2), as well as pupillary parameters, such as diameters and area. The time-stamped fixation events in the gaze data are assigned to their nearest visual focus (i.e., a colour, shape, or position). Fixations landing off the screen are ignored because they are not considered relating to the interaction with the system.

4.4.1 Synchronisation

The synchronisation between gaze and speech is an important prerequisite before performing further analysis. Accurate modern hardware clocks in the pcs allow synchronisation between modalities by having synchronous indicators (e.g., timestamps) in the signals [51] [272].

The eye tracker system used in this study (Eyelink II) outputs the fixation and saccade events, with each event having a start and an end time. The EyeLink eye tracker uses a saccade-picker approach to identify the saccades first based on the velocity, acceleration and distance moved between samples and then treats the rest of the data as fixations. This is quite different from another main event parsing approach, a fixation-picker approach,

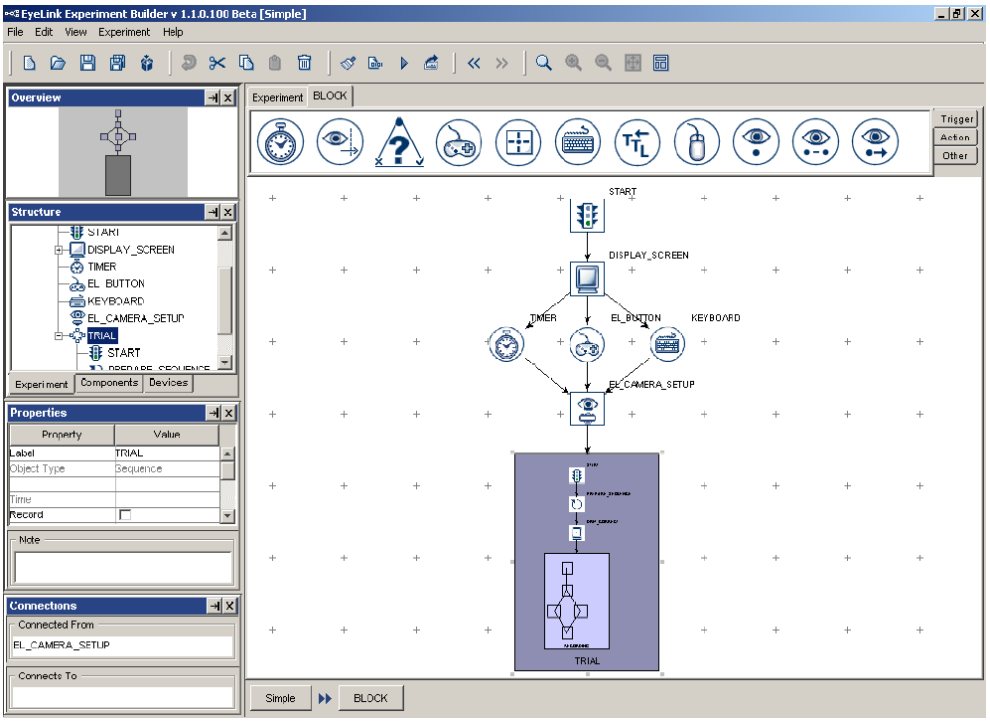


Figure 4.5: A typical example of the Trail (Session) Level design flow. The start of a trail can be triggered by keyboard, controller button, or a timer. The left-hand side panel lists the structures of the flow and the detailed properties of each component.

which has an assumption for the dispersion of fixation and minimum duration of fixation; once fixations are identified, the rest of the data segments are treated as saccades. The fixation-picker method is typically adopted by an eye tracker with a low-sampling rate because of the lack of the ability to do the velocity-based event parsing. The high-speed EyeLink eye trackers use the saccade-picker approach [268].

Speech and gaze data collection run in parallel on the Eyelink 2 host pc within a single application. The recording application's messaging handling allows the latency between speech and gaze to be less than the eye-tracker sampling period (2ms). This enables the synchronisation to be achieved by aligning the timestamps in gaze and speech signals. In addition, the result update on the screen controlled by the wizard is also time-stamped by the application using the same scheme.

```
0 52500000 sil -31282.957031 sil
52500000 54300000 sil -1400.342407 sil
54300000 55200000 r+eh -845.675964 red
55200000 55500000 r-eh+d -273.802216
55500000 56400000 eh-d -767.236267
56400000 66200000 sil -7957.309082 sil
66200000 67800000 s+er -1482.486938 circle
67800000 68100000 z-er+k -295.782257
68100000 68400000 er-k+l -290.114380
68400000 69300000 k-l -767.858276
69300000 69500000 sil -186.863464 sil
69500000 70200000 ax+n -673.877502 on
70200000 74100000 ax-n -3210.383545
74100000 80300000 sil -3937.406006 sil
80300000 80700000 ah+n -327.561432 one
80700000 81100000 ah-n -300.417419
81100000 93600000 sil -8218.645508 sil
```

Figure 4.6: An example of the phoneme-level time-aligned transcription. The first and second columns are start and end time (ns) respectively. The third and fourth columns are phoneme and its score. The last column is the word composed by the phonemes.

4.4.2 Speech transcription

Users' speech is recorded and stored in mono PCM WAV format. Recordings are time-aligned transcribed in three steps. The first step is by a human transcriber, and the second step uses an ASR system employing forced alignment [33] to segment words temporally with silences. The ASR was built using HTK [325], a popular software toolkit used in speech processing research (see section 2.3.1), and trained on the WSJCAM0 corpus of British English [271]. Further details of this ASR system may be found in Chapter 7. The last step is to manually validate the transcribed results by comparing them with the speech clips. An example of the phoneme-level transcription is shown in Figure 4.6.

4.4.3 Calibration errors and quality assessment

In order to know where the user's eyes are fixating on the computer screen, the system must learn what the eyes look like when they are fixating on the pre-known locations.

This supervised learning procedure is called calibration. Typically, it is followed by a validation procedure to estimate the confidence in the information learnt.

During the recording, the calibration may not remain reliable because the users may move unwillingly during the task, even though they are instructed to keep still. Also, the weight of EyeLink 2 eye-tracker is not ignorable. Wearing the head-mounted device may cause discomfort. This may lead to a reduction in calibration stability.

The calibration process assumes that the user always focuses at screen depth as all targets during training are presented on a computer screen. Therefore, any vergence movement (i.e., simultaneous movement of both eyes in opposite directions) may not be recognised correctly and may adversely affect the measured eye position. This also means that when fixation depth is altered, such as the user looking beyond the screen into the distance, the gaze position may not be accurate. During the experiment, this is mainly caused by the user being tired and losing focus. Thus, users are allowed to take a rest between each two recording sessions.

The recording application provides an effective means of objective quality assessment. Prior to each session recording, a drift correction is performed where the user fixates at a location to allow the eye tracker to correct the drift errors. The drift correction may fail, which leads the session to be discarded. The failure can be caused by un-correctable drift errors due to the head or eye-tracker movement. Figure 4.7 shows the examples of the good eye-camera positions and moved positions that are more likely to cause calibration errors.

In addition, each session recording is viewed thoroughly from beginning to end for the subjective quality assessment. The recording is played in three modes. First is the standard Animation View; the recording video is played back with the recognised gaze position overlaid on the task map. The Spatial Overlay View (Figure 4.8) allows looking at events by placing them where they were detected in space. The visual objects are also placed on the screen in the Overlay View. The larger the 'circle' is, the longer the gaze was fixed there. These circles can be seen as histograms with the visual objects being

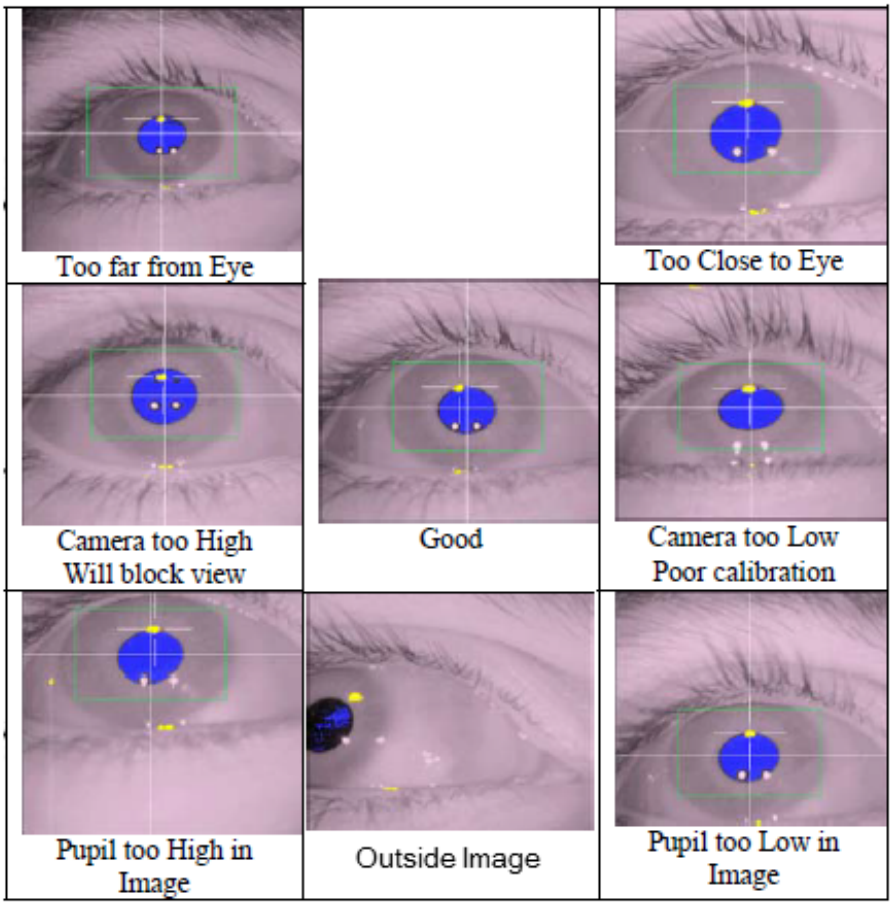


Figure 4.7: Examples of the good eye-camera positions and unsatisfactory positions that are more likely to cause calibration errors, from the EyelinkII manual.

the bins. If the majority of the fixations do not fall on any object, it can be reasonably assumed that there is a calibration error. The middle-left part shows the start time of each fixation event, and the bottom-left part shows more details, including the assigned nearest visual objects. By assessing the assigned objects, an effective subjective assessment can be done. Similar to the Spatial Overlay View, the Temporal Graph View (Figure 4.9) allows the user to view session data as a trace plot, where the horizontal axis represents time and the vertical axis represents the X and/or Y gaze location. The assessment can be performed by comparing the positions and the assigned visual objects.

The loss of calibration is more likely the longer the session. Users typically take longer to complete the task in acoustic noise. Also, the noise might amplify their discomfort. The calibration errors have caused some sessions to be rejected (see section 4.5.2), but

the rejection rate is acceptable and does not compromise the objectives of the corpus collecting.

4.5 Main Data Collection

4.5.1 Adding acoustic noise

The multi-speaker babble noise is used in the main corpus collection because it elicits greater behaviour change compared to white noise (stated by pilot study results, see section 5.3), and it better represents the real-world noisy environments. The noise from the NoiseX-92 corpus [308] is amplified to three noise loudness level. Thus, the following four noise level conditions are adopted:

1. No-noise condition as the baseline condition.
2. Light noise level - quieter than general speech in clean environment, average home environment.
3. Louder noise level - approximately conversation speech level at 1 meter.
4. Most noisy level - louder than speech, approximately outdoor commercial area.

The original noise has a mean sound pressure level (SPL) of 48.75dB [35], which is approximately average home environment level. Based on a previous work with the NoiseX-92 corpus [308] and this study's hypothesised real world environment for building a noise-robust HCI system (e.g., outdoor commercial areas), the noise is amplified by -6dB(N1), 6dB(N2) and 15dB(N3) respectively using the original noise as the reference. The resulting three levels of babble noises are 42.75dB (SPL), 54.75dB (SPL) and 63.75dB (SPL).

A participant begins his sessions from N0, processing to N3 and completes 5 sessions under each noise condition. These 20 sessions are called a session group. Details of the session structure will be described in the next section.

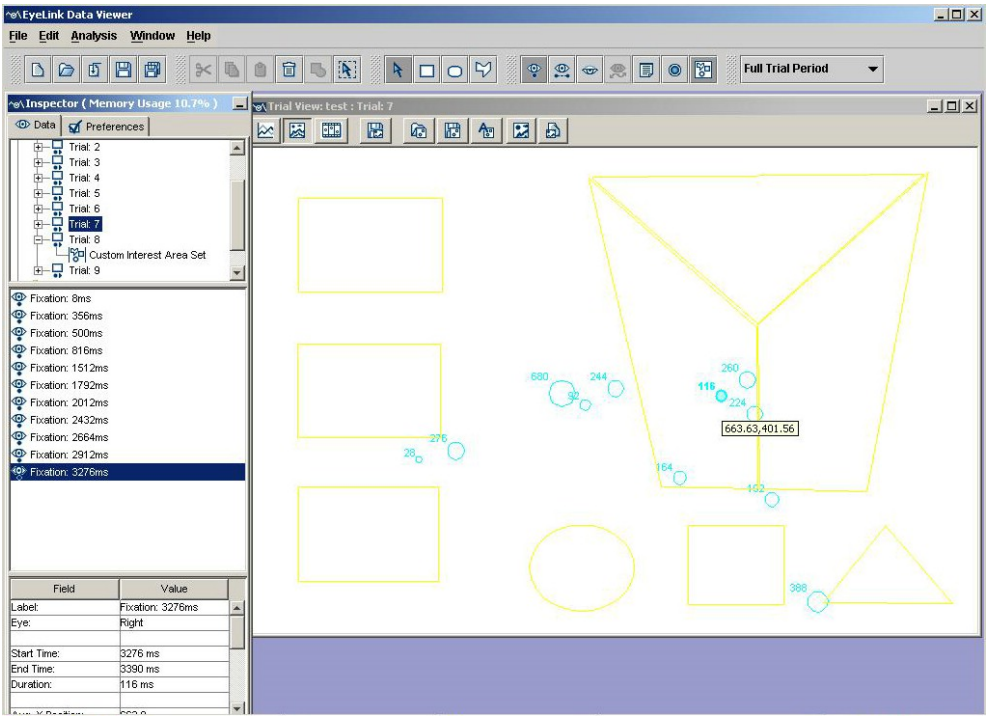


Figure 4.8: Spatial Overlay View interface (main windows inside Experiment Builder). The gaze position is overlaid with the visual objects on the screen for the ease of view.

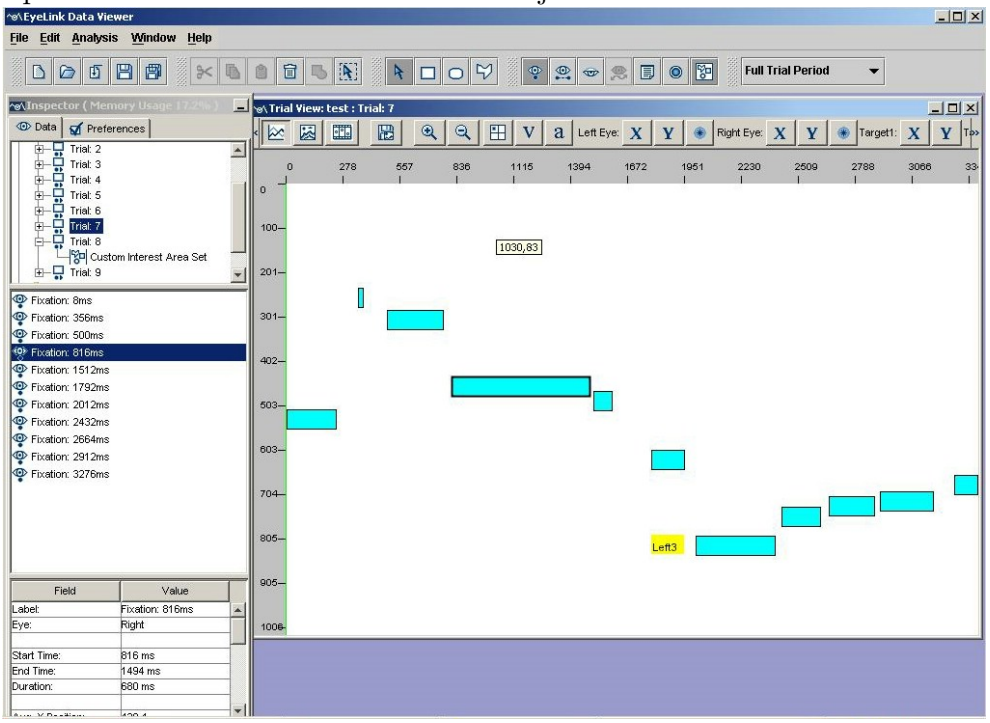


Figure 4.9: Temporal Overlay View interface (main windows inside Experiment Builder). The horizontal axis represents time and the vertical axis represents the X and/or Y gaze location.

4.5.2 Session structure

In total, 400 sessions are recorded. They are divided into 20 groups, and each contains 20 sessions. Within each group, sessions 1-5 are sessions with no noise (N0), sessions 6-10 Noise Condition 1 (N1), sessions 11-15 noise Condition 2 (N2), and sessions 16-20 Noise Condition 3 (N3). An eye-tracker re-calibration is conducted when changing noise condition. Each participant records 2-4 groups, and participants are given a break of approximately 10 minutes between groups to reduce any possible harm caused by the noise.

There is some data loss from calibration failures, such as unwilling movement or dislodging of the head-mounted apparatus as discussed in section 4.4.3. Consequently, 23 task recordings are discarded. This 6% loss is acceptable and expected in line with findings from experiences of other eye-tracking studies [321] [302]. Therefore, 377 valid sessions are collected. None of the discarded sessions are recorded in the no-noise condition. It is noticed that the number of sessions discarded has a positive relationship with the noise level added, which can be a clue of an increasing impact of the noises and longer session duration on calibration errors. Among the valid sessions, there are in total 11 sessions during which the wizard has misrecognised the user's instruction due to the acoustic noise. The overall misrecognition rate is 2.9%. A summary of the sessions collected is listed in Table 4.2.

4.6 Summary

This chapter has discussed Wizard of Oz simulation setting as a research method for data collection. The methodology implementing WoZ simulation and the system apparatus are described. The task is designed to encourage participants to use both gaze and speech modalities actively to interact with the system controlled by a wizard. The ES-N corpus data collected from the experiment is carefully processed for labelling and quality assessment. Certain restrictions still exist in the eye-tracker experiments. The calibration is not

		N0	N1	N2	N3
Noise Level		None	Noisex-92 Babble		
Noise Level (SPL)			42.75dB	54.75dB	63.75dB
Sessions	Participant				
	A	20	20	14	10
	B	10	10	10	10
	C	10	10	10	10
	D	15	15	15	14
	E	15	15	15	15
	F	15	15	14	10
	G	15	15	15	15
	Total	100	100	93	84
	Discarded	0	0	7	16
	Misrecognised	0	0	4	7
	Misrecognition rate	0	0	4%	8%
Duration (ms)	Mean	17.9	28.4	53.2	85.6
	Std	9.8	10.2	17.4	25.1

Table 4.2: A summary of the sessions collected. The noise level is measured by sound pressure level (SPL). There are more sessions discarded in noisier conditions. It typically takes longer to complete the task in acoustic noise.

perfectly reliable, and the causes are discussed. Some recording sessions are discarded, which is typical in eye-tracking experiments and in line with other studies.

A pilot data collection is conducted to justify the use of acoustic noise. The babble noise demonstrates the ability to better elicit the change of gaze behaviour and promote more intelligible communications. The detailed results of the pilot study will be presented in the next chapter, Chapter 5.

The main corpus collection experiment is performed with four noise level conditions and more participants. The main data collected will be further analysed in the next chapter with a 'gaze Lombard effect' revealed.

The MI approach outlined in section 3.4 will be applied in the next chapter for the selection of noise type using the pilot data and for the noise condition inference using the main data.

CHAPTER 5

ACOUSTIC NOISE INFERENCE AND THE GAZE LOMBARD EFFECT

For a gaze-contingent ASR system in which speech and gaze are used as inputs, modelling the relationship between these two modalities is a key requirement to achieve better recognition. When environmental acoustic noise is involved, the behaviour of each individual modality and their relationship may change accordingly. Therefore the dependency of the speech-gaze relation upon acoustic noise must be understood. In Chapter 4, the collection of the ES-N corpus was described. In this chapter, using the corpus data collected, the acoustic Lombard effect is validated and the 'gaze Lombard effect' revealed. The relationship between gaze and speech in acoustic noise are investigated.

Section 5.3 shows results of the pilot study conducted prior to the main data collection for investigating different noise types. In Section 5.4 and 5.5, the acoustic Lombard effect and the 'gaze Lombard effect' are explored respectively. In section 5.6, mutual information (MI) measures are applied for investigating the semantic and temporal relationship between gaze and speech in acoustic noise are used for the acoustic noise condition inference. Section 5.7 is the summary of the chapter contents.

5.1 Motivation

For the acoustic Lombard effect (see section 2.3.3 and 2.3.5), researchers have studied the effect in both speaker-independent and speaker-dependent experiments [148]. From the speaker-independent experiments, the general tendency of the change was studied, while in speaker-dependent experiments, the researchers seek the consistency that can be used for general modelling of the Lombard effect. However, the variability from speaker to speaker is significant and can be affected by gender, language, environment, and inherent personal strategy for adapting noise [144].

Moreover, the Lombard effect and its speaker-variability have been shown to present not only in speech but also in visual modalities, such as mouth, head, and jaw movement (see section 2.4.2). It is of the great interest to know whether gaze is also affected. In this chapter, the 'Lombard effect' in gaze and its relationship with speech are investigated, using both speaker-independent (aggregated) data and speaker-dependent (participant) data. Using the aggregated data, the general tendencies of gaze change in noisy environments are explored while variability is investigated across participant data.

As discussed in 2.3.4, knowledge of the acoustic noise condition could improve the ASR performance. For instance, an optimum acoustic model can be trained to match the noise condition. However, if the 'gaze Lombard effect' also varies widely across speakers like the acoustic Lombard effect, predicting noise using gaze features would not be desirable. Thus, a measurement of their relationship might perform better in noise-inference. For this purpose, a SVM classification task is performed to determine whether noise condition prediction using information from gaze is robust to between-person variation, compared to the case where the relationship with speech is considered.

5.2 Assumption of Normality

In statistics, performing parametric tests, such as t-tests relies upon the 'Assumption of Normality'. When comparing two sample means, the 'Assumption of Normality' can

be described as 'the sampling distribution of the mean is normal (Gaussian)'. Due to the Central Limit Theorem - 'given random and independent samples of N observations each, the distribution of sample means approaches normality as the size of N increases, regardless of the shape of the population distribution' - the distribution of sample means is approximately normal when the sample size is 'sufficiently large'. In practice, when performing a two-tailed t-test for comparing two sample means, the 'sufficiently large' sample size requires $N > 30$ [239].

In this thesis, the t-test is used with sample size $N > 30$ when comparing sample means unless specifically notified. For example, for 'gaze Lombard effect', the normality test and the corresponding non-parametric tests are performed, because it is considered adequate to reveal the distribution characteristics when reporting this novel observation statistically. However, using the t-test will not compromise the comparison results. All the significance values and other statistical tests in this study are performed using IBM SPSS.

5.3 Pilot Study

Before main data collection, a pilot study is performed to test the experiment settings, explore the data to collect and the relationship between gaze and speech, justify the acoustic noise type to use, and prepare the wizard (discussed in section 4.3.3). For these purposes, the aggregated data for all participants is used in the initial analysis to compare the impact of different noise types. Also it is of the interest to explore the idea of measuring the relationship between gaze and speech by applying the mutual information (MI) measurement (described in section 3.4).

5.3.1 Tests for noise type comparison

Babble noise and white noise are compared in terms of their effect on the gaze and its relationship with speech by measuring MI. The noises used are from the Noisex-92 corpus

[308]. During the experiment, the order of the noise types is randomised. Two common gaze metrics suggested by Jakob [137] are measured to indicate the gaze change: fixation duration and saccade length (see section 2.2.2).

MI is used to measure the dependency between speech and gaze. Refer to the expression for MI (expression 3.2), let g be the gaze focus and w be the spoken words; the MI between gaze events $G = (g_1, g_2, \ldots, g_n)$ and speech events $W = (w_1, w_2, \ldots, w_m)$ is written as:

$$I(G;W) = \sum_{g \in G, w \in W} p(G = g, W = w) \log_2 \frac{p(G = g, W = w)}{p(G = g)p(W = w)} \tag{5.1}$$

It is hypothesised that higher dependency between speech and gaze indicates a tighter coupling (e.g., the user uses gaze more actively to assist speech) leading to more intelligible communication. The noise type that elicits greater gaze change and promotes more intelligible communication will be selected to be used in the main data collection experiment (see section 4.5).

More details of the MI measure can be found in section 3.4 and 5.6.

In addition, a usability test regarding efficiency (time to finish the task), accuracy (whether the user's instruction is understood and properly updated on the screen), and the feeling of being able to interact naturally is performed by questionnaire.

5.3.2 Pilot data results

The fixation durations and saccade lengths of 4 participants are compared under three noise conditions: no-noise, white noise, and babble noise. Two acoustic noises used are of the same original loudness level (SPL) (see section 4.5.1). The results are listed in Table 5.1.

From Figure 5.1, it is shown that, in babble noise, a greater magnitude of change is observed in terms of longer fixation duration ($p < 0.05$) and shorter saccade length ($p < 0.001$). The MI results in Figure 5.3 indicate higher dependency ($0.34bit$ $p < 0.05$)

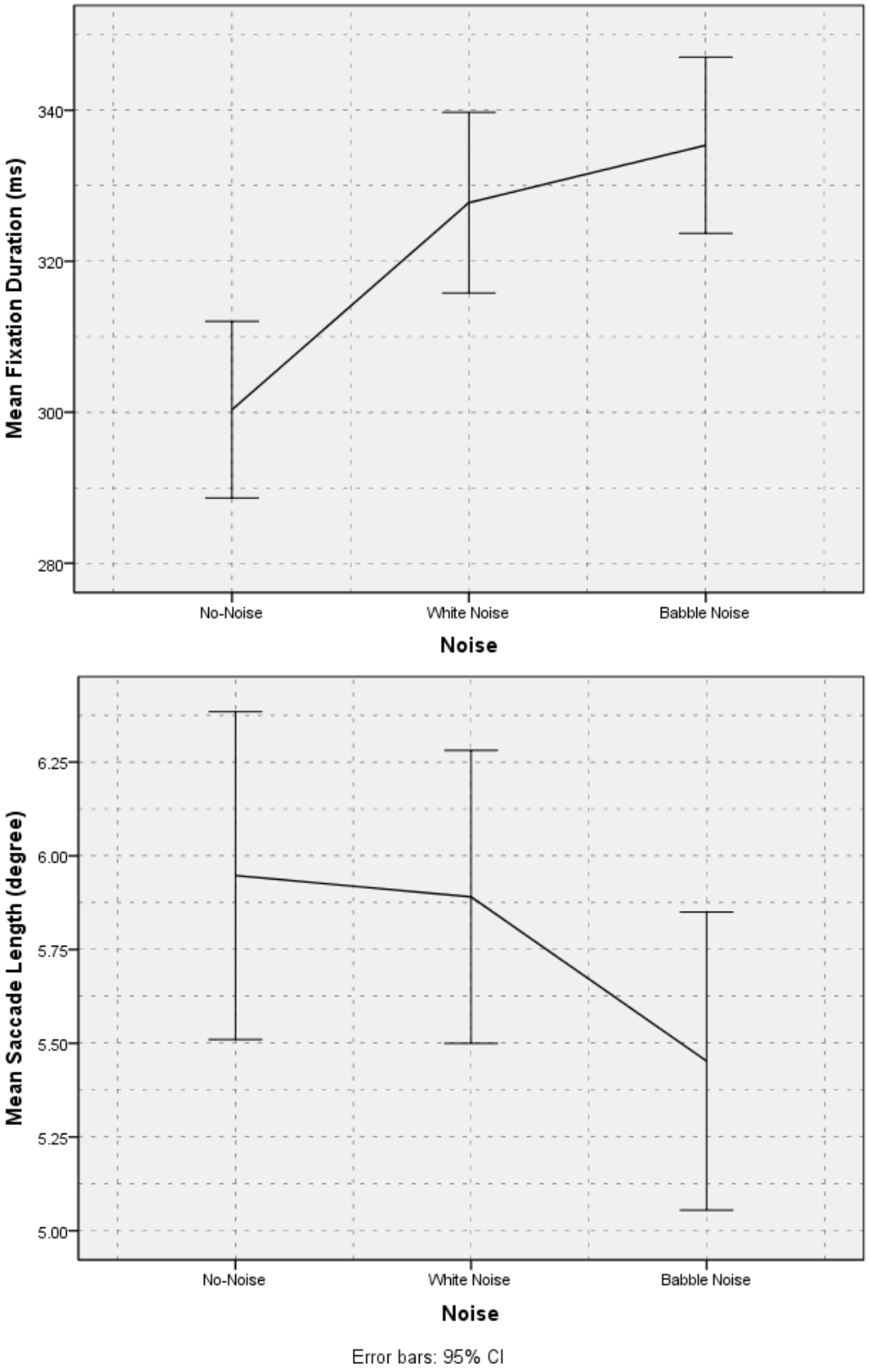


Figure 5.1: Participants' fixation duration and saccade length with 95% confidence interval error bars in no-noise, white noise, and babble noise conditions. Longer fixation duration and shorter saccade length is observed in babble noise condition.

	Fixation Duration		Saccade Length	
	Mean	St.d	Mean	St.d
No-Noise	300.34	227.1	5.94	5.45
White Noise	327.76	230.4	5.89	5.76
Babble Noise	335.57	261.7	5.45	5.83

Table 5.1: Fixation Duration and Saccade Length across three noise types. A greater magnitude of change is observed in terms of longer fixation duration and shorter saccade length in babble noise compared to white noise and no noise. the order of the noise types played is randomised.

between gaze and speech in babble noise.

For the completeness of gaze features analysis, it is tested that whether the fixation duration and saccade length in different noise types are normally distributed. The Kolmogorov-Smirnov test (adapting Lilliefors significance for normal distribution [180]) and the Shapiro-Wilk test [286] are applied for the data normality test, as well as the normality Q-Q plot (Figure 5.2). In the Q-Q plot, the data does not hug the normality line tightly, indicating that the distribution is not normal (Gaussian). Failing the normality test ($p < 0.001$ in both tests), the non-parametric Kruskal-Wallis test [29] is used as the significance test with the null hypothesis that the data comes from the same distribution.

Together with the previous Lombard effect studies discussed in section 2.3.3, it can be concluded that babble noise not only elicits a stronger Lombard effect for speech, but also a greater change in gaze characteristics. One plausible explanation for the higher dependency is that participants use gaze more to aid the speech for a better instruction. This supports the previous findings [146] [148] [58] that speakers' behaviours, both vocal and non-vocal, produced in babble noise are proved to be more intelligible than those produced in no-noise condition, while it is the opposite in white noise. The results make babble noise a more desirable acoustic noise type to use in the main corpus data collection (see section 4.5.1) for being able to elicit more intelligible gaze and speech behaviours.

The usability test shows that most session durations are between 1 and 2 minutes. By the end of all sessions, the users are satisfied with the results updated by the system (controlled by the wizard), and they do not feel restricted in the communication style.

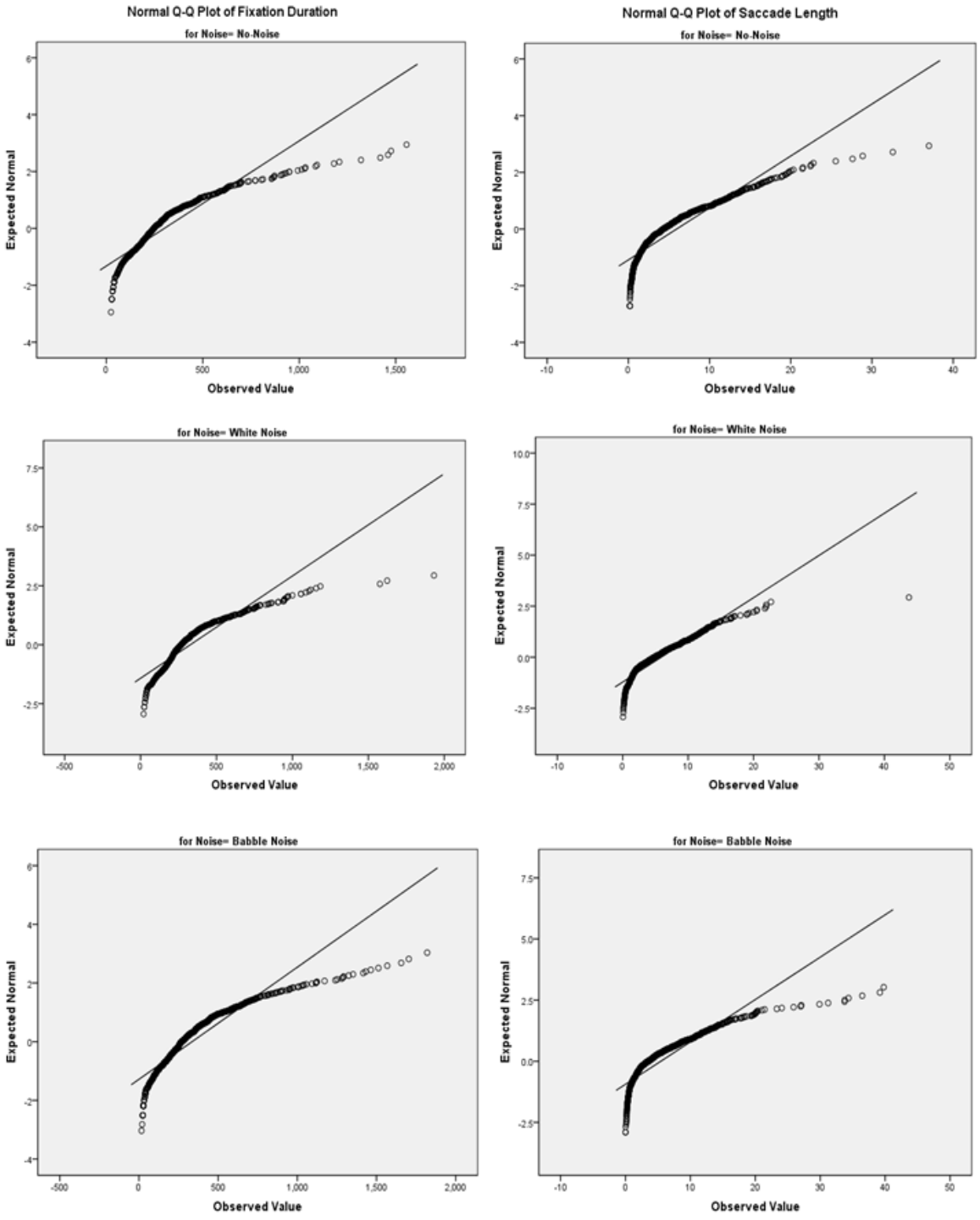


Figure 5.2: The normality QQ-plot for fixation duration (left column) and saccade length (right column) in no-noise (top row), white noise (middle row), and babble noise (bottom row).

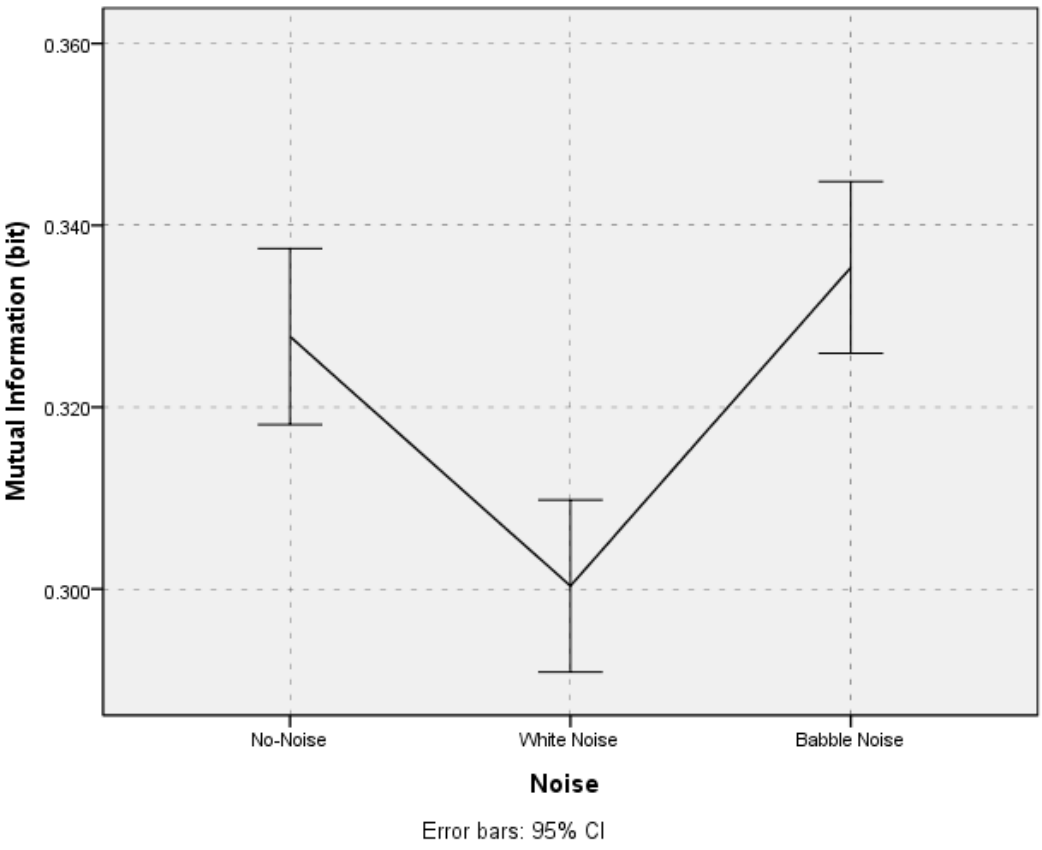


Figure 5.3: The MI results (with 95% confidence interval error bars) between speech and gaze across three noise conditions. (The theoretical maximum value of the MI $I_m \approx 3.20$.)

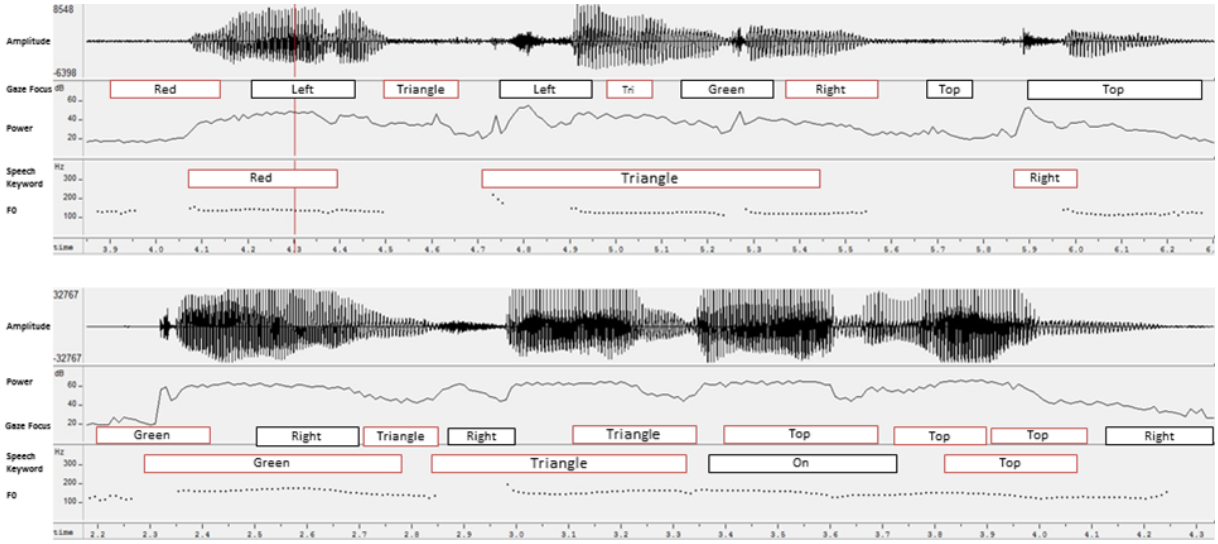


Figure 5.4: An example of captured gaze and speech features without (top) and with (bottom) environmental acoustic noise. The labelled spoken words and visual attention foci are overlaid.

5.4 Acoustic Lombard Effect Analysis

In section 2.3.3, the acoustic Lombard effect is discussed. In this section, this phenomenon is validated with the ES-N corpus data.

5.4.1 Speech data test performed

To study the dependency of the gaze-speech relationship upon acoustic noise, it is helpful to investigate the change of speech (i.e., acoustic Lombard Effect) first. Suggested by the previous researches [148] [204], a test is performed to reveal the change of speech rate (words per minute), fundamental frequency (F0), and the average speech power across four different noise level conditions using the main data collected. These parameters are extracted using an open source tool for sound visualization and manipulation. Figure 5.4 illustrates an intuitive example for captured gaze and speech features in no-noise (top) and noisy (bottom) environments. In each environment, the top view shows the amplitude, centre view the gaze foci and speech power, and bottom view the words spoken and f0.

Speech characteristic	Additive Noise Level			
	N0	N1	N2	N3
wpm				
μ	62.5	63.3	59.6	**58.5**
σ	18.6	13.4	13.4	**12.9**
p value	-	0.96	0.40	**0.01**
F0				
μ	137.5	141.7	**153.3**	**162.9**
σ	30.9	29.9	**30.7**	**32.1**
p value	-	0.13	**0.00**	**0.00**
Power				
μ	30.9	29.9	**30.7**	**32.1**
σ	2.4	2.3	**1.8**	**2.1**
p value	-	0.13	**0.00**	**0.00**
n	100	100	93	84

Table 5.2: Speech characteristics in 4 noise conditions. Bold indicates significance compared to no-noise condition. The n is the number of samples.

5.4.2 Results

As a result of adding noise, a significant ($p = 0.01$) decrease in the average speech rate across all 377 recorded trials is observed (Table 5.2), from 62.5 words per minute (wpm) in no noise to 58.5 wpm for 15dB noise (row 1). Likewise, the average pitch across all participants (F0) significantly increases ($p = 0.00$) from 137.5 to 162.9 (row 2). The power also significantly ($p = 0.00$) increases from 58.8 to 62.6. The variance in wpm shows a greater change, from 18.6 to 12.9, as noise increases compared to the other measures. These are aggregate measurements for all participants.

The changes in speech characteristics on an individual basis revealed that pitch (F0) and power increase significantly for all users. The change in wpm (Table 5.3), however, is observed significant in participants A, E, and G, and shows both an increase and decrease in average wpm. Interestingly, in all cases, the variance is observed to fall in noise. The results support previous results regarding the Lombard effect, that the adaptation varies across the speaker with power and F0 being the main changes in speech [148].

Participant	$\delta\mu$	$\delta\sigma$	p value	n
A	**-21%**	**-35%**	**0.01**	**30**
B	-12%	-49%	0.63	20
C	+21%	-12%	0.47	20
D	+02%	-64%	0.99	29
E	**+38%**	**-44%**	**0.00**	**30**
F	-14%	-40%	0.35	25
G	**-36%**	**-44%**	**0.00**	**30**

Table 5.3: Breakdown in the average speech rate (wpm) for the 7 participants. The percentage is the relative change from N0 to N3.

5.5 Gaze Lombard Effect Analysis

In this section, the main gaze data is analysed, and a 'gaze Lombard effect' is revealed.

5.5.1 Gaze data test performed

The gaze recording data contains all the fixations and the in-between saccade events from the 377 valid sessions. In this study, semantically related events in gaze and speech are the visual attentions related to the spoken words: for example, the fixation on 'circle' and the word 'circle'.

In section 5.3, the pilot study reveals that the 'gaze Lombard effect' involves changes in fixation duration and saccade length. In this section, the 'gaze Lombard effect' is further explored by investigating the change of fixation duration and saccade length in acoustic noise with the speaker-independent data. However, as a participant's visual attention information is contained in the fixation events, more detailed analysis of fixations are reported compared to saccade length with the between-people variability. The study also addresses the question of whether the changes in fixation duration across noise conditions come from the fixations during silence or during speech and their relative comparison (i.e., the 'gaze Lombard effect' may differ depending on whether someone is speaking). If the changes are more significant during speech, then the finding is likely to be more useful for automatic speech recognition.

A number of tests are performed to answer the following questions:

- Does a person's fixation duration and saccade length distribution depend upon acoustic noise? How is this change related to the noise level? What is the nature of the distribution?
- Is there a distinction between the fixations during speech and the fixation during silence? Is this distinction dependent upon acoustic noise?
- How does the fixation duration distribution during speech and during silence change respectively in acoustic noise?
- How does the change of fixation duration distribution vary between different people?

5.5.2 Results

Overall fixation durations and saccade lengths

The histogram of the fixations under the 4 noise conditions is shown in Figure 5.5. In each noise condition, a certain level of skewness (a measure of the distribution asymmetry) is evident because the histogram does not fit the normal distribution line exactly. The normal distribution lines in 4 noise conditions show a decreasing kurtosis (a measure of whether the distribution is peaked or flat relative to a normal distribution) between $100ms$ and $300ms$, which can be a sign of the increasing amount of longer fixations.

Further tests for the normality assumption are performed. Both the Kolmogorov-Smirnov test (adapting Lilliefors significance for normal distribution [180]) and the Shapiro-Wilk test [286] are applied to the fixation data. Table 5.4 shows that by both tests, the null hypothesis that the mean distribution is normal is rejected ($p < 0.001$) in all 4 noise conditions. From the normality Q-Q nlot (Figure 5.6, 5.7), it can be explicitly read that the distribution under each noise condition does not hug the expected normal line tightly. The results also (as the pilot study results in section 5.3) suggest that the use of a normal distribution to model fixation duration is not desirable. The result of a Box-Cox transformation [278] with $\lambda = 0$ (i.e., a log transformation) indicates the distribution is

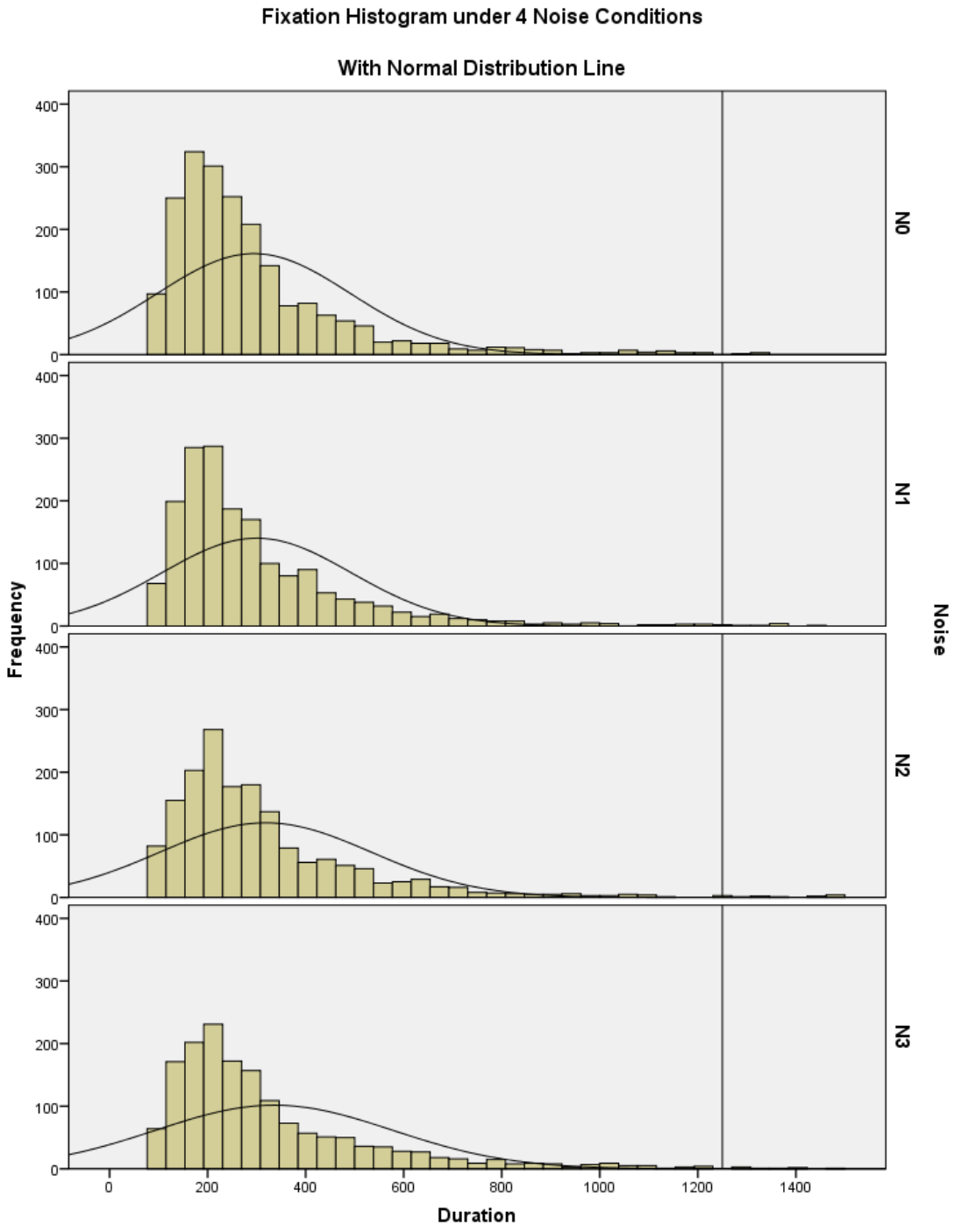


Figure 5.5: The histogram of fixations under 4 noise conditions and the normal distribution lines with a decreasing kurtosis across the noise conditions (all sessions and users). Compared to the normal distribution line overlaid, it is shown that the assumption of normal distribution is not evident.

	Noise	Mean	Std	Number of fixations
Duration	N0	294.21	196.66	2067
	N1	299.83	193.269	1767
	N2	318.14	215.86	1677
	N3	335.55	241.871	1602

Table 5.4: The number of the fixations with the means and the standard deviations across 4 condition groups N0, N1, N2 and N3. The Kolmogorov-Smirnov test and the Shapiro-Wilk test indicate all results are significant ($p < 0.001$).

log-normal ($p > 0.05$, for more discussions see section 6.3.3 and Figure 6.3).

A test is taken to compare the fixation durations of 4 noise conditions. The non-parametric k-independent Kruskal-Wallis test [29] is used to compare the fixation distributions under 4 noise conditions. The null hypothesis that these 4 groups of fixations come from the same distribution is rejected ($p < 0.001$). However, the test does not indicate whether pairs come from the same distribution (there are 6 possible combinations for creating pairs from the pool of 4 conditions). Because it is of interest to find out whether the variation in fixation duration is introduced related to acoustic noise, the fixations under N0 (no-noise) are compared to the fixations from the other conditions (N1, N2, and N3). Therefore the Kruskal-Wallis test is repeated between N0&N1, N0&N2, and N0&N3. In the second and third combinations, the null hypothesis is rejected, which means the difference is significant to say that the fixations from N0 are different from the fixations from N2 ($p < 0.001$), and N3 ($p < 0.001$). The results suggest that noise modifies fixation duration. Figure 5.8 illustrates the increasing trend of the fixation mean value when noise increases. While in the first combination (N0&N1), the null hypothesis cannot be rejected ($p = 0.202$), which means the fixations under N0 are likely to come from the same distribution of fixations under N1. This supports the finding that the increase of fixation duration is related to the increase of noise loudness level; the duration hardly differs when there is only little noise difference.

The same statistical tests are performed on the saccade lengths. Similarly, failing the normality test ($p < 0.05$), the Kruskal-Wallis test suggests that the saccade lengths under N0 are significantly different from those under N2 ($p < 0.001$) and N3 ($p < 0.001$), while

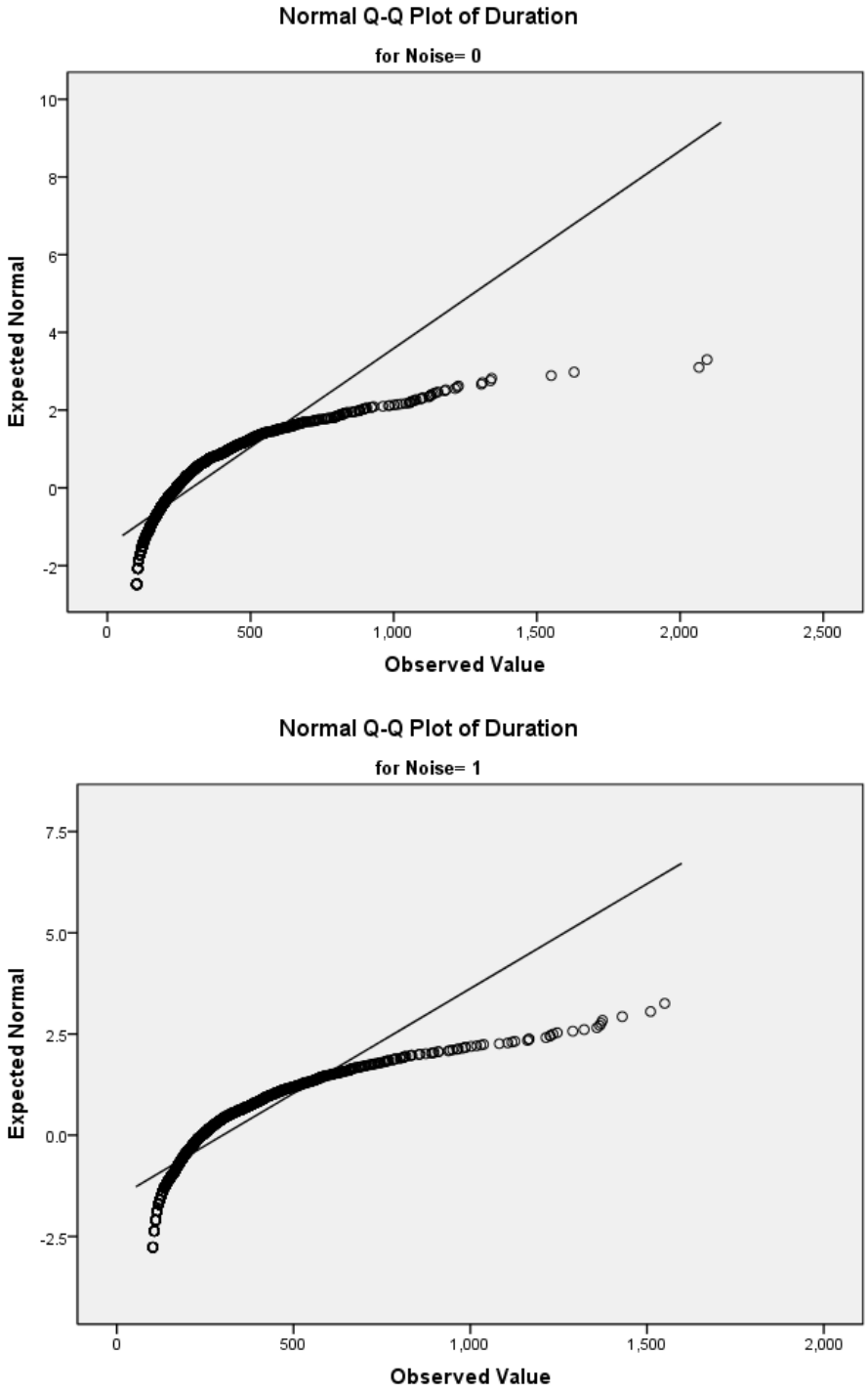


Figure 5.6: The normality Q-Q plot of fixations under N0 and N1. From the figures, it is evident that the mean distribution is not normal.

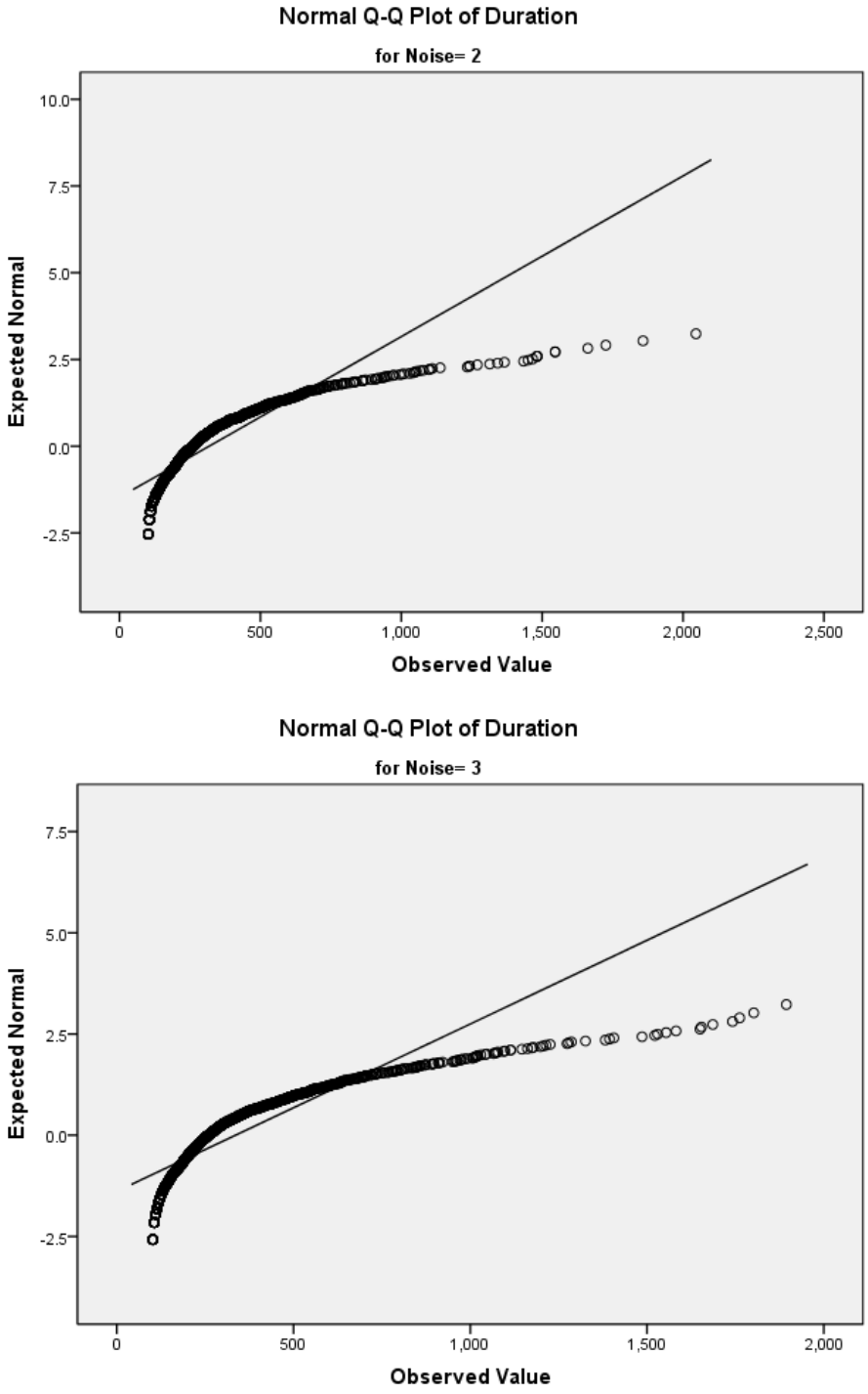


Figure 5.7: The normality Q-Q plot of fixations under N2 and N3. From the figures, it is evident that the mean distribution is not normal.

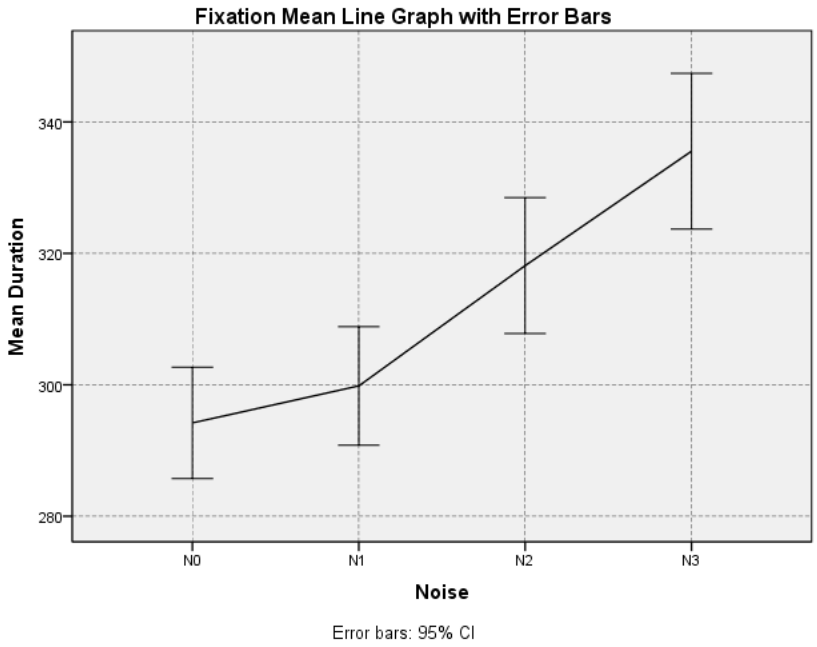


Figure 5.8: The mean value of the fixation durations (ms) across 4 noise groups with error bars of 95% confidence interval. An increasing trend is evident. Together with the results of the Kruskal-Wallis test, it can be confirmed that the increase of noise loudness level introduces significantly longer fixations.

no significant difference is found in those under N1 ($p > 0.05$). This finding agrees with the fixation result, that the change is dependent upon noise level. However, contrarily a decreasing trend is noticed (see Figure 5.9).

Fixation durations related to speech occurrences

In order to discover whether the fixation behaviour is different while the participant is talking, the fixations are divided into two groups that are 'During Speech' and 'During Silence'. A fixation is considered 'during speech' if it has temporal overlap with a speech event (a piece of temporal segment starting with the onset of a word's occurrence and ending with this word's termination). Otherwise, it is considered 'during silence'. There are 4 noise conditions, so in total, the fixations are divided into 8 groups (see Table 5.5). It is hypothesised that the fixation duration while someone is talking is different (e.g., longer) than when not.

From Table 5.5, an increasing trend of the fixations count percentage during speech

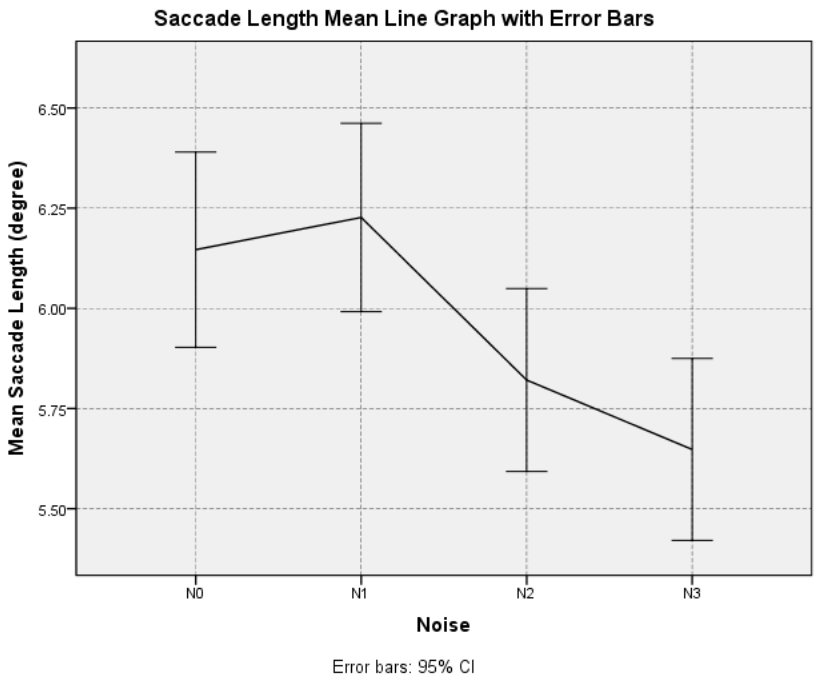


Figure 5.9: The mean value of the saccade length across 4 noise groups with error bars of 95% confidence interval. A decreasing trend is evident. Together with the results of the Kruskal-Wallis test, it can be confirmed that the increase of noise loudness level introduces significantly shorter saccades.

		Number of fixations		Percent	
Duration	Noise	Speech	Silence	Speech	Silence
	N0	659	1408	31.9%	68.1%
	N1	577	1190	32.7%	67.3%
	N2	597	1080	35.6%	64.4%
	N3	649	953	40.5%	59.5%

Table 5.5: The frequency statistics of the fixation events 'during speech' and 'during silence'. An increased percentage of the fixation count 'during speech' is observed.

		Noise			
		N0	N1	N2	N3
Mean(ms)	During Speech	327.80	318.65	360.08	387.97
	During Silence	278.48	290.71	294.96	299.86
Increase from N0	During Speech		−2.8%	9.8%	18.4%
	During Silence		4.1%	5.9%	7.7%

Table 5.6: The mean value of the fixation duration 'during speech' and 'during silence' across 4 noise conditions. The increase level 'during speech' is more notable comparing with the level 'during silence'.

(from 31.9% to 40.5%) is observed when acoustic noise level rises. Within each noise condition, the results in Table 5.6 show that there is a distinction between the two groups. The mean fixation duration 'during speech' is longer than that 'during silence' ($p < 0.001$, $p = 0.006$, $p < 0.001$, and $p < 0.001$ for N0, N1, N2, and N3 respectively).

A test is performed to compare the fixations 'during speech' in no-noise condition (N0) with those in acoustically noisy conditions (N1, N2, and N3). The results suggests that when the light noise (N1) is added, there is no difference in terms of the mean fixation duration 'during speech' ($p = 0.706$). However, as the loudness level keeps rising, there starts to be an increase of duration ($p = 0.028$ for N2). When the noise is even louder, a more notable increase of the mean fixation duration 'during speech' is observed ($p < 0.001$). The increase amount is $32.28ms$ between N2&N0 and $60.17ms$ between N3&N0.

Similarly, the comparison of the fixations 'during silence' in no-noise condition (N0) with those in acoustically noisy condition (N1, N2, and N3) shows no significant difference in the light noise (N1, $p = 0.077$). However, as the noise loudness level rises, increased mean fixation durations are observed ($p = 0.012$ for N2 and $p < 0.001$ for N3). The

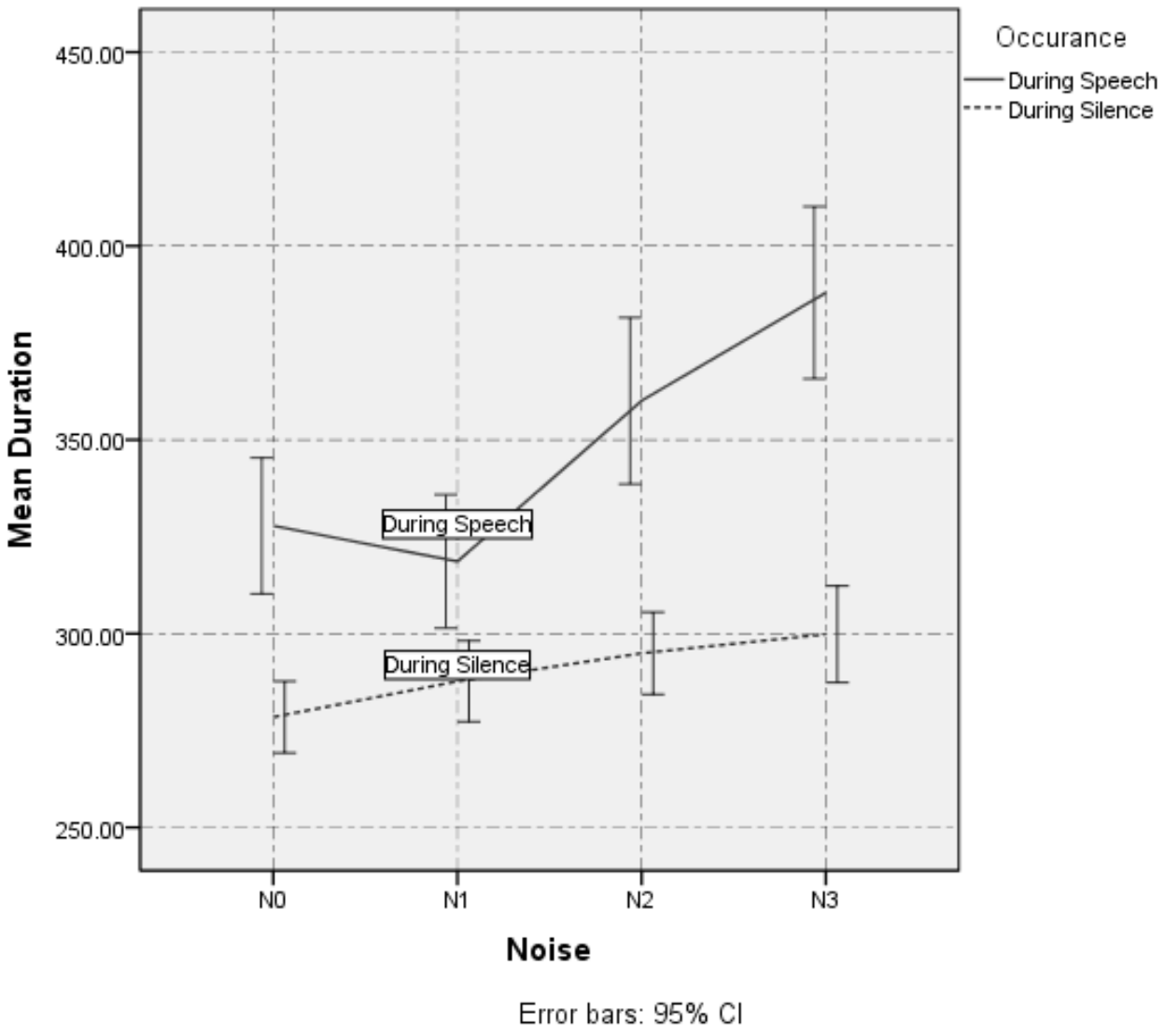


Figure 5.10: The distinction between the fixations 'during silence' and 'during speech' across 4 noise conditions with error bars of 95% confidence interval. An enlarging trend of such distinction can be observed from the figure.

increase amount is 16.48ms between N2&N0 and 21.38ms between N3&N0.

It has been illustrated that mean fixation durations during both speech and silence are correlated to the noise level. By comparing the increase level of 'during silence' with that of 'during speech' (Table 5.6), it can be shown that when noise loudness level rises, the increase of the fixation duration 'during speech' is more notable than that 'during silence', resulting in an enlarged distinction between these two groups within a noise condition. Figure 5.10 illustrates this conclusion.

Between-person variation

To analyse the variability of the fixation duration changes in noise across different people, the Kruskal-Wallis test is applied to compare the change of the mean value between N0 and N3 for each participant. Table 5.7 shows that, although an average increase can be observed from the data, some of them (B,D,E) are not statistically significant but no decrease is observed. The similar variation also exists in the vocal adaptation (acoustic Lombard effect), as described in section 5.4. This suggests that each individual responds differently to noise. It may be argued that in an acoustically noisy environment, a person might choose to adjust another modality (a rising of speech power or f0) rather than change gaze. A pilot test is performed to explore this potential relationship between fixation and vocal adaptation. The results in Table 5.7 show that the three least significant participants rank as the first three in speech power increase, while the correlations for F0 and wpm are not evident. However, more data and more specific experiments would be required to take this further.

Summary of findings

Longer fixations and shorter saccades are observed when the noise loudness level increases (Figure 5.8 and 5.9). The change has a correlation with the noise loudness level. For a specific noise condition, a person has a fixation duration distribution with higher mean value during speech than when not talking. It has been demonstrated that this distinction

Participant	$\Delta fixationduration$	$pvalue$	$\Delta power$
A	+17%	0.050	+4.8%
B	+3%	0.401	+18.9%
C	+17%	0.038	+1.2%
D	+0.4%	0.723	+5.3%
E	+12%	0.195	+4.4%
F	+16%	0.043	+2.8%
G	+26%	0.000	+3.3%

Table 5.7: Breakdown in the relative mean fixation duration change from N0 to N3 for the 7 participants, showing that 3 participants' lengthening is not significant. The three least significant participants (highlighted) rank as the first three in speech power increase

is lengthened as the noise loudness level increases (Figure 5.10).

It can be concluded that in the acoustically noisy environment, besides adjusting speech behaviour (the acoustic Lombard effect), gaze behaviour also changes (the 'gaze Lombard effect') depending on whether someone is speaking. Both the speech and gaze change are subject to a variation between different persons. It needs to be noted that the gaze behaviour change might also be related with the wizard's misrecognition of user instructions. However, as the misrecognition rate is very low (2.9%, see section 4.5.2), it is considered that the results are not compromised by the fact.

5.6 Acoustic Noise Inference

In section 5.4 and section 5.5, the acoustic Lombard effect and the 'gaze Lombard effect' are revealed to vary between persons. To compare whether the measurement of gaze-speech relationship has less between-person variation and performs better in acoustic noise inference (discussed in section 5.1), their relationship and the dependency on acoustic noise are investigated in this section.

The relationship between gaze and speech is defined as their semantic and temporal relationship (see section 3.3). This is not the same as the previous section where gaze 'during speech' and 'during silence' were compared or the same as correlation between F0, power, and fixation duration. Here it is assumed that speech and gaze are information

events and the use of mutual information (MI, see section 3.4) is proposed to be the measure of quantifiable relationship strength.

5.6.1 Density estimation in MI calculation

In section 2.2.6, the benefit of using preceding and co-occurring gaze events when combining with speech is discussed. For the inference of acoustic noise using MI-based measures, two corresponding relationships (see section 3.2) between gaze and speech - where people look at objects prior to naming them - Object Naming (ON) and where people look at objects during naming them - Mediating Attention (MA) - are considered. Both roles can be considered cognition and lacking direct measurability. In both roles, the events are defined as objects viewed and spoken words. However, coupling functions may be defined differently for MI calculation. With MI value calculated for the two gaze roles, a feature vector can be formed and used to infer the acoustic noise condition, which is a measurable variable. For example, one may expect the MI for the cognition role of 'mediated attention' to increase relative to 'object naming' if there is a significant amount of noise in the room, and this relative difference can be utilised by a classifier for the noise inference.

A general coupling framework between modalities and a MI-based approach for modelling the relationship were described in section 3.3 and 3.4. Refer to expression 3.1:

$$f_r(G, W) = h(f_r^s(G, W), f_r^t(G, W)) \tag{5.2}$$

Figure 5.11 illustrates the two different measures of calculating mutual information. For object naming (Figure 5.11 lower image), the *a priori* information about this role from cognitive psychology is that speakers fixate on objects between 740ms [197] and 932ms [105] before naming them for a duration on average of 600ms [104] - the 'eye/voice span'. Thus, couples between gaze and speech events g and w are weighted based on these temporal and semantic constraints $f_r(g, w) = r_{g,w}$ so that $0 \leq r_{g,w} \leq 1$. Referring to the

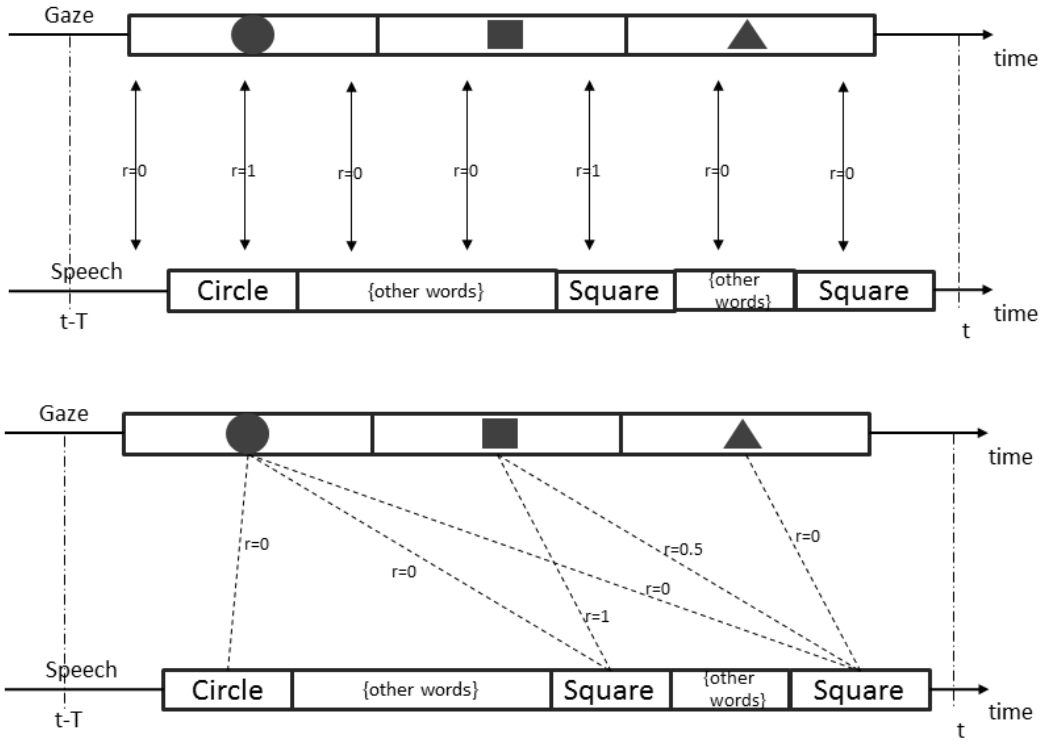


Figure 5.11: Example of calculating two different measures of mutual information (upper one for mediating attention and bottom one for object naming) between gaze sequence G and speech sequence W for noise-inference. Density estimation is based on frequentist estimates of the observed couples and their strength r, which is determined from temporal and semantic constraints defined in the multimodal coupling function $f_r(G, W) = r$ (expression 5.2).

example in Figure 5.11, looking at an image of a square 800ms prior to saying 'square' will be represented by a couple with a higher value ($r_{g,w} = 1$) than a couple representing looking at it 1500ms prior($r_{g,w} = 0.5$) whereas looking at the triangle and saying square has $r_{g,w} = 0$.

For mediated attention (Figure 5.11 upper image), it is defined as when one person gestures with their eyes towards a visual focus to guide the attention of another. This 'intention to guide' reflects cognition, and once again, cannot be directly validated. The joint density in MI represents the co-occurrence of objects viewed and words spoken, is based on how much they overlap in time. Consequently, unlike object naming where couples are defined between events that occur asynchronously, couples are made periodically. The value of the coupling function $r_{g,w}$ is based entirely on the semantic relatedness of the events g and w and not temporal constraints, simplifying expression 5.2.

For both gaze roles, the joint density for event pairs seen in couples is estimated based on weighted frequentist estimates of the event couples:

$$p(g, w) = \overline{r_{g,w}} \frac{N_{g,w}}{N_{seen}} \tag{5.3}$$

where $N_{g,w}$ is the number of couples between gaze event g and speech event w, and N_{seen} is the total number of couples observed. $r_{g,w}$ is the weight of a event pair and equivalently $\overline{r_{g,w}}$ denotes the averaged value here. Within any temporal window T over which MI is calculated, there may be events that are observed only in another window (e.g., other objects and words). Therefore, to preserve the axiomatic assumption of unit measure, the joint density for events seen in the data, but unseen as event couples in window T, are uniformly estimated from the probability mass that is not assigned to the seen joint probabilities in expression 5.3:

$$p_u(g,w) = \frac{1 - \sum_{g,w \in GW_{seen}} p(g,w)}{N_{unseen}} \tag{5.4}$$

where N_{unseen} denotes the number of unobserved event pairs. Expressions 5.3 and 5.4 ensure that if a gaze role is not prevalent, then the joint density tends towards a uniform distribution, and thus, results in a low value of MI (i.e., independence). In contrast, if a gaze role is prevalent, then the joint density will be less uniform, resulting in higher values for MI. Because the constituents of the joint density are observed events, the MI for different temporal windows becomes comparable regardless of its constituents. Marginal densities are calculated from the joint density.

The overall MI measure considering all event types in case and another modality can be considered in matrix form whose dimensions are equal to the number of types of events seen in each modality (i.e., number of objects and object names):

$$I(G;W) = [i(g;w)]_{g=1\ldots n_g, w=1\ldots n_w} \tag{5.5}$$

In this example, n_g and n_w are the number of objects and words respectively, i the individual MI calculation of a (g,w) pair.

With values for $I(G;W)$ calculated for the two gaze roles, a feature vector can be formed and used to infer the acoustic noise condition.

5.6.2 Test conducted

For two hypothesised relationships between spoken words and gaze events (ON and MA), MI is calculated to measure the relationship strength. The calculation is repeated for four different noise conditions (N0, N1, N2, and N3), and the results are compared.

In section 5.4 and 5.5, the speech and gaze characteristics (e.g., wpm, f0, fixation dura-

tion) are shown to vary between people. This makes these individual measures less useful in a classifier, which uses these features in estimating the noise level in the environment. To evaluate whether the MI measure offers more value in discriminating between noise conditions and is more consistent between people, classification tests have been performed to compare the MI measure with speech and gaze features. For the purpose of inferring acoustic noise condition, a support vector machine (SVM) classifier is built.

As a summary statistics, the MI value needs time to be stabilised. Commonly, the MI is calculated on the entire data sequence [51] [273] [275]. Therefore, for the demonstration, a long fixed time window is used for each speech sentence to be within the time window - the entire sentence is used for the MI calculation.

To formalise:

- Test 1: Calculation of MI when gaze is used for mediate attention
- Test 2: Calculation of MI when gaze is used for object naming.
- Test 3: A classification test for the acoustic noise condition using MI, gaze, and speech features to compare the discriminability. The feature with higher accuracy is more favourable.
- Test 4: A classification test for the participants using MI, gaze, and speech features to compare the variability. The feature with lower accuracy is more favourable (i.e. the feature does not vary between people).

5.6.3 MI results (Test 1 & 2)

The MI is calculated over the entire temporal window across 377 sessions. The results in Figure 5.12 reveal that for the MI measure based on the coupling of mediating attention, it significantly increases ($p < 0.01$) with noise: from 0.33 bits for no noise (N0) to 0.42 bits for the noisiest (N3) condition. Meanwhile, there is no significant increase ($p > 0.05$) for the MI measure that is based on the coupling of object naming. This suggests that

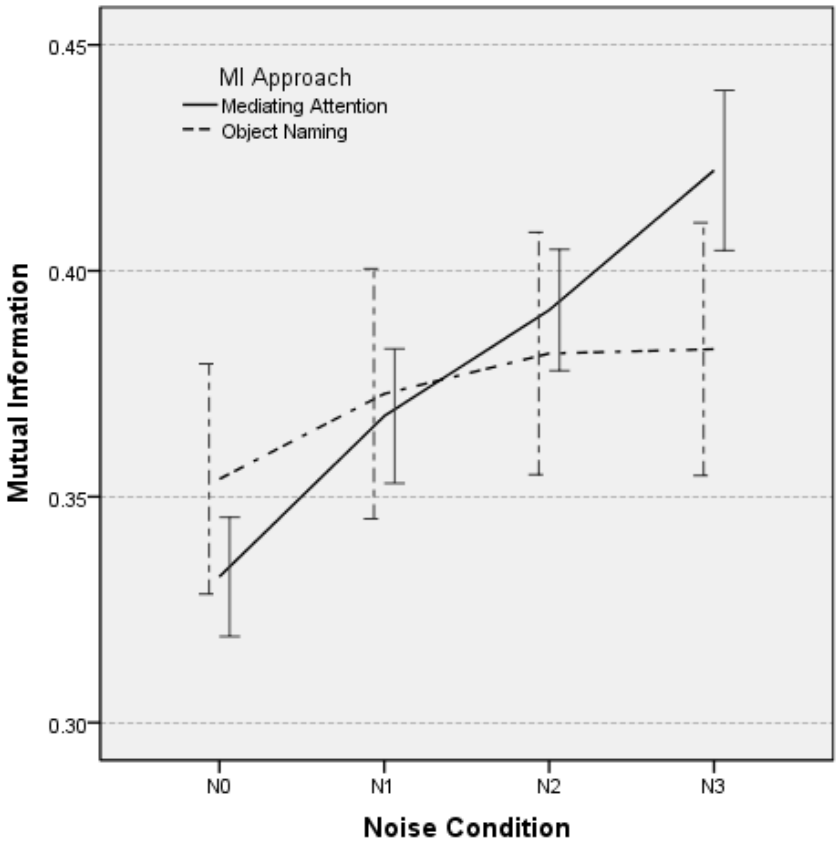
this predefined coupling (MA) is more prominent in noise, suggesting that the users use their gaze (as instructed to in the WoZ descriptors section 4.3.2) to assist the wizard to understand instructions.

5.6.4 Discussions as to SVM experimental setup

SVM is a popular classification technique based on the concept of defining optimal (maximum margin) separating hyper-planes in the feature space when a kernel is used [52]. It provides a good trade-off between performance and computational complexity. Because of its easy implementation, task-independence, and generally satisfactory performance, SVM is used in the study for the evaluations of the MI measure.

Multi-class classification

Although SVM is considered the best off-the-shelf classifier for binary classification[21], extending it for multiclass classification is still an on-going research topic. The common approaches include 'one-against-all', 'one-against-one', and directed acyclic graph SVM (DAGSVM). The 'one-against-all' method is the earliest used approach [27]. In a k-class situation, k SVMs are constructed with each considering one class as positive and the rest negative. The classification of a new instance is done according to the maximum output among all SVMs. The 'one-against-one' method introduced by Knerr [165] is also known as 'pairwise coupling' which constructs one SVM for each class pair, resulting $k(k-1)/2$ SVMs in total. The classification of a new instance is done by a 'max-wins' voting strategy, in which the class with the most votes determines the instance classification. The earliest use of this implementation can be seen in the studies of Friedman [90] and Kressel [169]. The third approach DAGSVM is proposed in a study of Weston [319]. Its training phase is the same as the 'one-against-one' method, but in the testing phase, a binary directed acyclic graph, which has k leaves and $k(k-1)/2$ nodes, is used. Each node is a binary SVM; therefore, the classification process involves going through a path before reaching a



	Noise							
	N0		N1		N2		N3	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
MI1(MA)	0.332	0.091	0.368	0.103	0.391	0.089	0.422	0.112
MI2(ON)	0.354	0.176	0.373	0.191	0.382	0.178	0.383	0.176

Figure 5.12: The MI values based on the coupling of mediating attention (MA) and object naming (ON) respectively. There is a significant increase for the first role when noise increases and no significant change for the second. (The theoretical maximum value of the MI $I_m \approx 3.20$.)

leaf node, which represents the predicted class. It is not easy to simply state which method is the best approach, but Hsu [123] demonstrates that the 'one-against-one' and DAGSVM are more suitable for practicality and ease of use. Therefore in this study, considering the relatively small feature space (MI, gaze, and speech) and moderate instance number (377 tasks), the 'one-against-one' method is selected for its easy implementation.

Kernels and parameters

Kernel functions allow SVMs to map the data into a different space where the maximum-margin can be achieved by using a hyper-plane. There are four kernel functions that have been found to work well in a variety of applications and they are linear, polynomial, radial basis function (RBF), and sigmoid. Among them, RBF kernel can handle the nonlinear relationship between class labels and features by nonlinear mapping. It has been demonstrated that the RBF kernel can behave like linear kernel [156] or sigmoid kernel [181] by adjusting the penalty parameter C and kernel parameter γ. It has fewer hyper-parameters than polynomial kernel [122] and less numerical complexity. Therefore, in general, the RBF kernel is a reasonable first choice [122] and it is selected for the classification tests here.

As mentioned above, there are two parameters for an SVM with an RBF kernel: C and γ. While the cost parameter C in SVM is the trade-off between training error and overfitting, γ is the parameter in RBF kernal:

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2} \tag{5.6}$$

where feature vectors are represented by the 'comparison function' K of two inputs x_i and x_j [312].

To identify a good (C, γ) pair so that the best classification performance can be achieved, a grid-search of two parameters is conducted. The grid search exhaustively looks for an optimal combination of C and γ within a chosen range that produces the

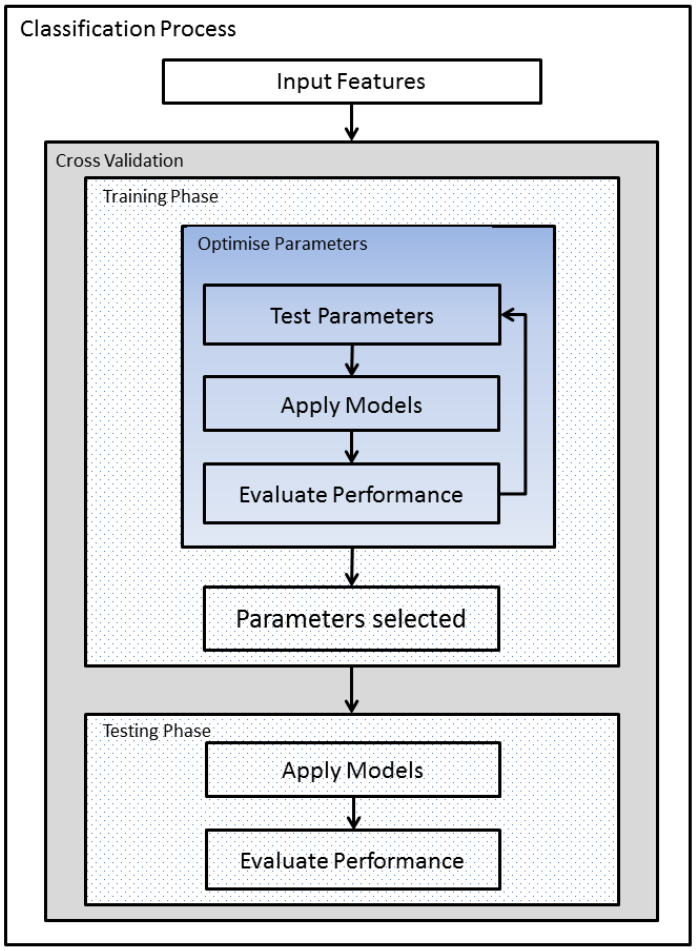


Figure 5.13: The classification process framework. The parameter optimisation is part of the testing phase.

best classification performance. Then this optimal combination of parameters will be used for constructing the SVMs in the testing phase. The framework of the whole classification process is illustrated in Figure 5.13.

Performing a grid-search with exponentially growing sequences of C and γ using cross-validation has been proved to be a practical method to identify good parameters [122] (e.g., trying $2^{-5}, 2^{-3}, \ldots, 2^{15}, \ldots$). Hence, a 7-fold cross-validation is employed in which the data is divided into 7 subsets with each subset contains data of one participant. Sequentially, 6 subsets are used in each iteration for training, and the remaining 1 subset is used for testing. The remaining components of the system in this paper are also evaluated using the same scheme.Using the cross-validation procedure also helps to reduce

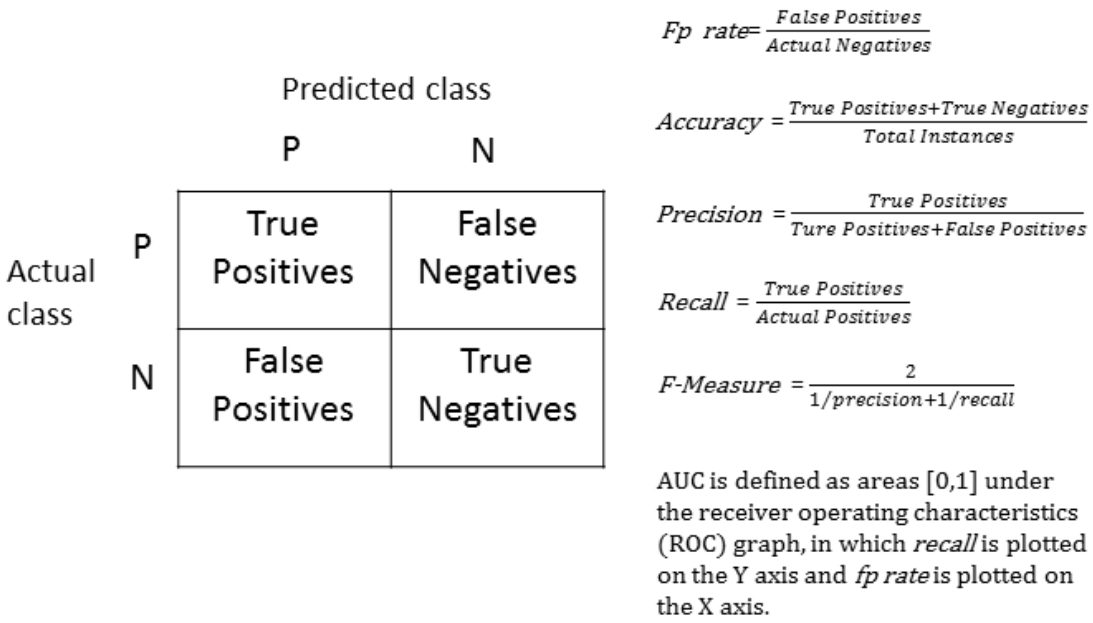


Figure 5.14: The definition of common classification evaluation metrics [82].

the over-fitting problem [200]. Typically, accuracy is used to evaluate classification performance. However, the limitations are mentioned by Provost [255] where the data is class-unbalanced. Due to the little class skew in the test data, accuracy is used for evaluating in grid-searches. Other metrics like precision, recall, F-measure, and area under curve (AUC) are also compared in the testing phase. The definitions of these metrics are illustrated in Figure 5.14. It needs to be noted here that the main objective is to compare the MI measure with other gaze and speech features under similar or identical settings rather than to build the best classifiers for inference task.

5.6.5 Noise classification results (Test 3 & 4)

During the classification tests, the gaze features (G), including fixation duration and saccade length; the speech features (S), including power, F0, and wpm; and the MI measure (MI) are tested as the input features respectively. The averaged values are used as that the estimation of MI requires the calculation of joint probabilities over a certain

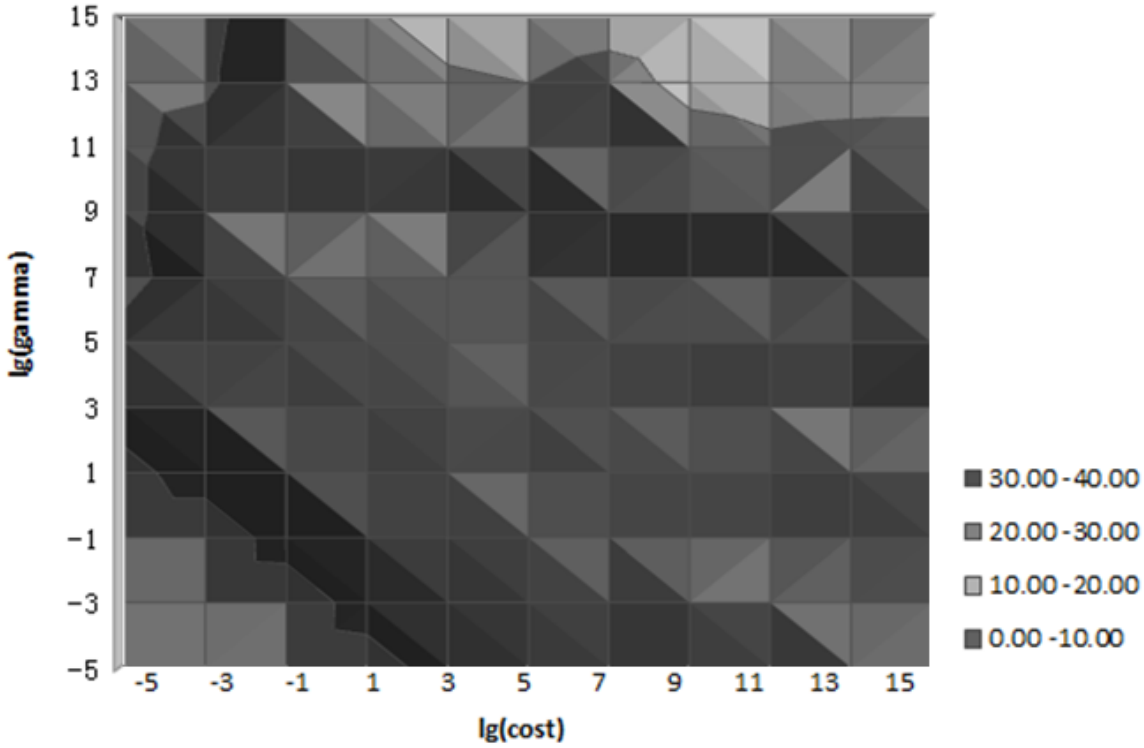


Figure 5.15: An example contour plot of classification accuracy. A grid search on $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ is performed.

period. The SVM parameter pairs (C, γ) are determined empirically using grid-searches. For example, a grid search of (C, γ) ranges from 2^{-5} to 2^{15} is conducted (Figure 5.15) for input feature MI, and the parameters $C = 2^5$ and $\gamma = 2^9$ are selected with a classification accuracy of 36.1%.

The SVM described in section 5.6.4 is trained to infer four noise conditions, and the results show classification accuracies ranging from 26.6% to 36.1% (Table 5.8), depending on the feature scheme for inputs. MI performed better than G (26.6%) or S (28.7%). Together with other evaluation metrics demonstrating that using MI bettered others in discriminating between noise conditions. The metrics reported are the weighted average over each target class.

The 36% classification performance for MI is 11% better than chance accuracy. To understand this issue better, inspecting the classifier confusion matrix (Table 5.9) indicates that misclassification is more likely to be made to adjacent (in terms of noise level) classes. Consequently, because the classifier works best at classifying data in the no-noise

Input Features	Acc	Precision	Recall	F-Measure	Avg AUC
MI	0.361	0.366	0.361	0.346	0.569
S	0.287	0.280	0.287	0.269	0.521
G	0.266	0.143	0.266	0.185	0.51

Table 5.8: SVM classifier performance for the four noise conditions (n=377). Feature scheme MI performs favourably compared to G and S respectively.

	Predicted Class			
	no noise	-6dB	6dB	15dB
no noise	61	25	10	4
-6dB	31	36	23	10
6dB	23	40	25	5
15dB	19	29	22	14

Table 5.9: Confusion matrix of instance numbers for noise classification results given in table 5.8. Misclassifications are more likely to happen between adjacent noise conditions (closer noise level gap).

condition and the noisiest condition, a 2-class SVM is trained and evaluated for classifying N0 and N3. This fact meets the finding in the previous section that the difference of MI is related to the noise level increase and that difference between N0 and N3 is statistically significant (see section 5.5).

The results from the two-class SVM in table 5.10 show that MI feature performs best at 71.2% - 21.2% above the chance accuracy of 50% (compared to 11% above the chance accuracy of 25% previously). The results support the findings that as acoustic noise increases, the change of gaze and speech characteristics vary between people, making MI a better measure to discriminate the noise level.

While MI has shown potential in discriminating within people due to the dependency of speech-gaze relationship upon acoustic noise in the environment, the results for the SVM classification task for discriminating 7 participants (Table 5.11) indicate that MI

Input Features	Acc	Precision	Recall	F-Measure	Avg AUC
MI	0.712	0.712	0.712	0.710	0.705
S	0.520	0.499	0.520	0.487	0.509
G	0.540	0.289	0.540	0.381	0.50

Table 5.10: Two-class SVM classifier performance for no noise and 15dB noise (n=184).

Input Features	Acc	Precision	Recall	F-Measure	Avg AUC
S	0.429	0.433	0.43	0.427	0.663
G	0.321	0.309	0.321	0.307	0.6
MI	0.18	0.227	0.18	0.115	0.509

Table 5.11: SVM classifier performance for 7 participants over all tasks (n=377).

performs worse than the G and S measures at 4% above chance accuracy compared to the best performing measures at 28% above chance accuracy. In this instance, however, worse performance is preferable because it indicates that MI is more robust to variation between people.

5.7 Summary

In this chapter, the relationship between gaze and speech is explored and the dependency upon acoustic noise is investigated. The motivation is to infer the acoustic noise condition in an ASR system for acoustic adaptation and to define the 'gaze Lombard effect'.

In section 5.4, tests are performed on the speech data, and the results are found to support the previous studies of Lombard effect, with fundamental frequency and power characteristics being the main adaption. In section 5.5, tests are performed on the gaze data. The results show that people tend to lengthen their fixation duration in acoustic noise, and more when speaking compared to when not. A decreasing tendency of the saccade length is also observed. Similar to the speech adaptation, the amount of this gaze adaptation varies between different people. However, no adverse adaptation (shorter fixation duration or longer saccade length) is observed. Some pilot arguments and tests are made regarding this variation in multimodal systems, but more work should be done in the future to support the finding.

After investigating speech and gaze independently, the semantic and temporal relationship between them and the change in noise is quantified using MI-based measures in section 5.6. The MI is used to be an indication of a relationship strength between gaze and speech. Considering the temporal asynchrony, two cognition gaze roles - object

naming (gaze precedes speech) and mediating attention (gaze co-occurs with speech) are compared. A feature vector is formed from these MI measures and used by a classifier to infer the acoustic noise condition.

The results demonstrate the potential for MI as a desirable measure for the acoustic noise inference for acoustic adaptation because it has less variation between people compared to other gaze or speech characteristics.

In Chapter 6, for a gaze-contingent ASR system, an event-based visual attention inferring framework based on interaction and environment change reaction gaze roles for language model adaptation will be described and evaluated. The findings in this chapter and Chapter 6 will be applied to an ASR system in Chapter 7 for the validation of ASR performance improvement (see Figure 1.1 in Chapter 1).

CHAPTER 6

VISUAL ATTENTION INFERENCE

In Chapter 5, the noise dependent speech and gaze relationship is explored with quantifiable measures, including mutual information, enabling the inference of the noise condition from gaze behaviours. This can be used to adapt the acoustic model in a gaze-contingent ASR. As the formalism is mentioned in the thesis earlier (Chapter 3), the relationship between gaze and speech is explored from the objective of determining the types of visual attention. In a previous research, Bednarik [21] was managed to infer gaze events related to issuing a command during an HCI task. In this chapter, the idea is developed further to consider the multimodal relationship with speech and different acoustically noisy environments. As words can be associated with visual attention, its inference allows language model adaptation in gaze-contingent ASR. In this chapter, a visual attention inference (VAI) framework is proposed and validated using the ES-N corpus.

6.1 VAI Implementation for the ES-N Task

In section 3.5, a formalism for VAI is described considering information events in gaze itself and their relationship to those in another modality. In this section, the general formalism for VAI is applied specifically to the ES-N corpus task in Chapter 4. A user speaks to a system that responds by (optionally) verbally confirming understanding and changing its display accordingly.

In the task, a user might speak the object name while looking at it. The system recognises a user's gaze (g_t) and speech events (w_t). Referring to the taxonomy (section 3.2, this interactional behaviour infers a task-oriented visual attention (TOVA). Consider a new object appearing on a display and gaze moving towards it. This is the gaze role of reactive visual attention (RVA) - i.e., a display change (d_t) and gaze orientating towards it. Considering a VA is associated with a gaze event sequence containing consecutive fixations and saccade, Figure 6.1 illustrates the applied taxonomy in VAI. In addition to TOVA and RVA, the task-independent visual attention (TIVA) may be assumed in the absence of TOVA and RVA. As discussed in the taxonomy section 3.2, the SVA is not considered in the scenario as there is only one user, and no social interaction (e.g., establishing agency or regulating interaction) assumed by the system.

Modalities in this application in addition to gaze are the user's speech and the system response - i.e., confirmation of a command issued and an update to the display. As discussed in section 3.6, expression 3.4 can be generalised as: $P(r_t = r|g_t = g, M) \propto P(g_t, M|r_t = r)P(r_t = r)$. Here $M = (W, D)$ is exploited, where W is the sequence of speech events and D the system response events, so that w_t and d_t are the speech and system response events at time t respectively. The multimodal coupling of a gaze event with a speech or a system response event $f_r(.)$ is determined by their semantic relatedness $f_r^s(.)$ and temporal distance $f_r^t(.)$ (see Figure 6.2).

Therefore, referring to expression 3.5, for TOVA (denoted by T) it is written as:

$$P(r_t = T|g_t = g, M) \propto P(g_t = g|r_t = T)P(M|r_t = T, g_t = g)P(r_t = T) \tag{6.1}$$

The multimodal coupling function for speech $f_r(g_t, W)$ and system response $f_r(g_t, D)$ are estimated respectively. The estimation of $f_r(g_t, W)$ is dependent upon whether the user's speech event w_v is task related (e.g., a command) and its temporal distance to g_t:

$$f_r(g_t, W) = \max_v f_r(g_t, w_v) \tag{6.2}$$

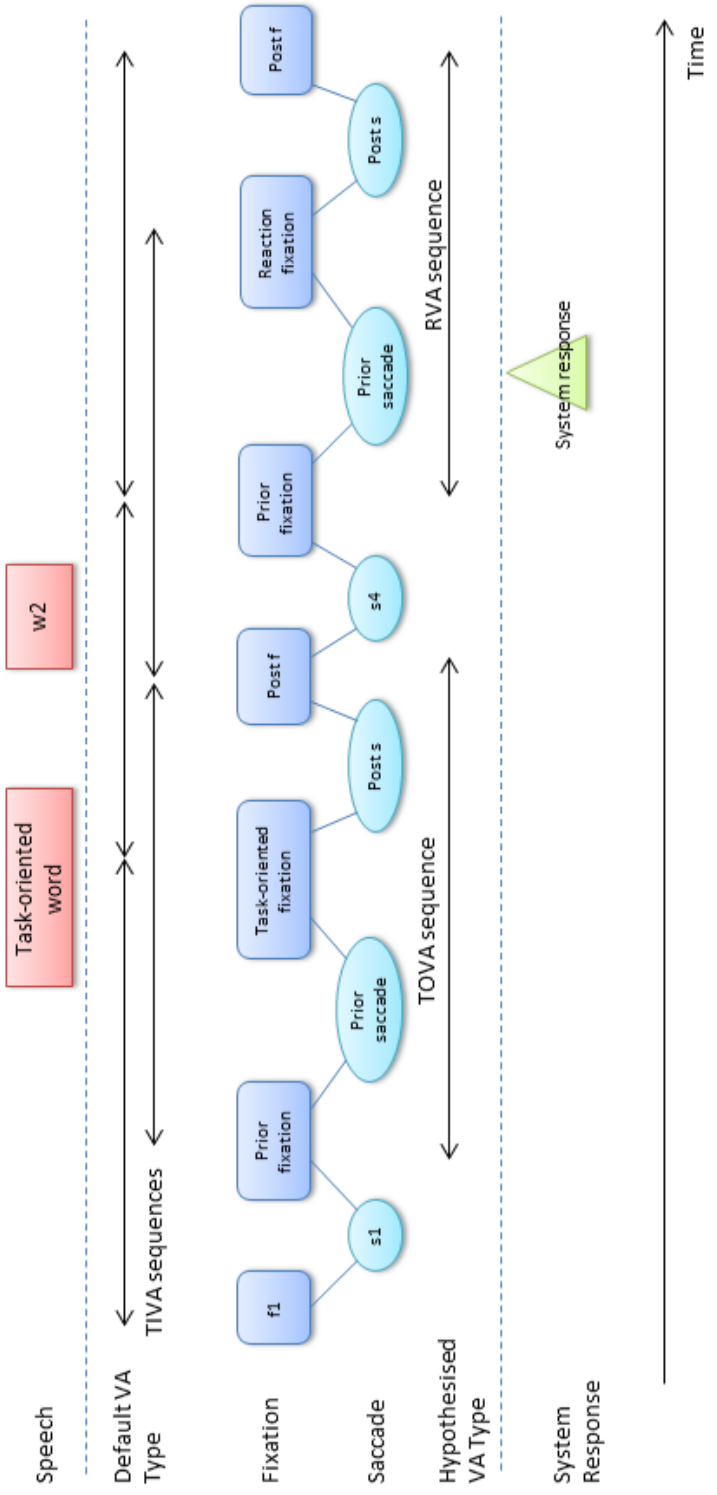


Figure 6.1: TIVA, TOVA, and RVA events in the form of sequences for feature extraction. Sequences are formed using a three-fixation window and a one-fixation interval. Each sequence contains three fixations and two saccade movements.

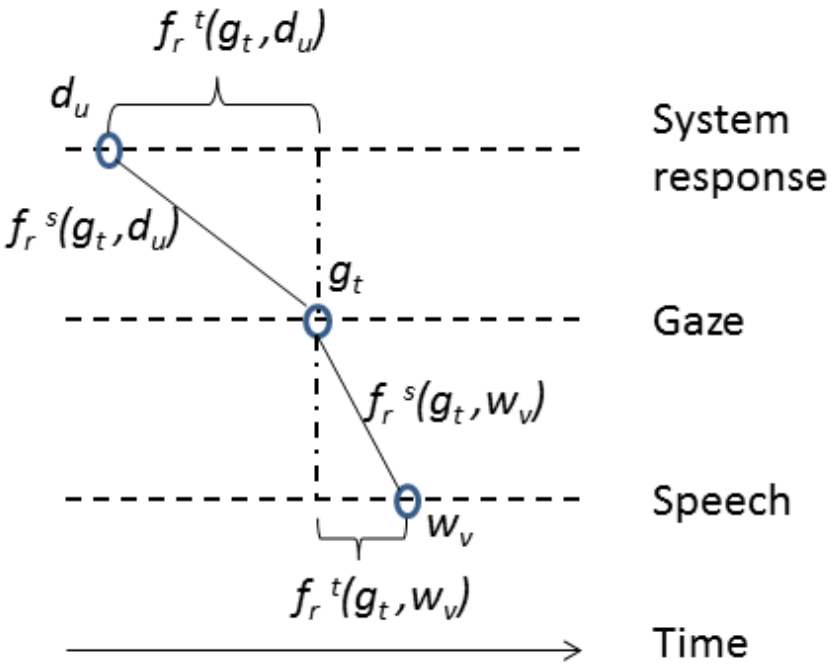


Figure 6.2: The coupling function between a gaze event g_t and a speech event w_v is determined by a semantic component f_r^s and a temporal component f_r^t.

where $w_v \in W$. The semantic component of the multimodal coupling function $f_r^s(.)$ (refer to expression 3.1) assigns a value of zero or one depending on whether the speech event w_v contains related dialogue:

$$f_r^s(g_t, w_v) = \begin{cases} 1 & \text{if } w_v = \text{task-oriented dialogue} \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

The temporal component of the coupling function gives greater weight to events nearer the time of the gaze event. For this, an exponential decay function may suffice:

$$f_r^t(g_t, w_v) = e^{-\Lambda_d |t-v|} \tag{6.4}$$

Thus, the multimodal coupling function in expression 3.1 is a product of the component temporal and semantic functions:

$$f_r(g_t, w_v) = f_r^s(g_t, w_v) f_r^t(g_t, w_v) \tag{6.5}$$

The product of two functions is employed so either function serves as a weight to the relationship.

Similarly for system response, $f_r(g_t, D)$ is defined. The key difference is that the coupling function is concerned with the (visual) system response D rather than speech W, therefore:

$$f_r^s(g_t, d_v) = \begin{cases} 1 & \text{if } d_v = \text{update to display} \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

$$f_r^t(g_t, d_v) = e^{-\Lambda|t-v|} \tag{6.7}$$

Inference of visual attentions thus becomes learning the parameters of the coupling functions and estimation of the densities that can be achieved by supervised learning.

6.2 Method

In the following sections, the evaluation for the taxonomy and inference framework discussed is undertaken employing the E-SN corpus data collected in the task described in Chapter 4. The coupling functions in section 6.1 are applied.

According to the taxonomy proposed, the gaze event in a standard human-computer-interaction (HCI) task can be distinguished as:

- TOVA, the events orienting to the task(s) assumed by the system function, which can be, for example, moving a block, pressing a button, or selecting an item.
- RVA, the events reacting towards the system responses which can be, for example, fixating on the pop-up information or confirming the block movement.
- TIVA, the events irrelevant to the on-going task, such as confusion, searching or looking away, as gaze is 'always on'.

The idea that gaze roles can be inferred from gaze characteristics alone (i.e. expression 3.7) is explored and compared with the addition of coupling function. To achieve this,

the densities $p(g_t|r_t = TOVA)$, $p(g_t|r_t = TIVA)$, and $p(g_t|r_t = RVA)$ are estimated in order to calculate $P(g_t = g|r_t = r)$ for each gaze event. The gaze features to use for VAI are assessed in section 6.3.5 based on the performance of role discriminability.

With densities estimated, their change as a result of adding noise to the environment is analysed. This is for three reasons. First, it lends support to the task assumed by the system to determine whether someone is looking at spoken objects. Second, it provides an insight into how gaze behaviour (and thus class density estimation for VAI) changes as a result of environment factors - i.e., environment CA. Then, it exams the robustness of the framework to assist integration between gaze and speech in the noisy environment to be used in ASR.

The task assumed by the system (TOVA) is determined by its function. In this evaluation, this function (ASR) is posited to be appropriate integration with speech using the related information in gaze (e.g., to enhance ASR performance). A gaze-contingent ASR task does not force the user to use gaze deliberately (see section 2.2.1) and could be more valuable in acoustic noise where the recognition of speech modality becomes more difficult and less reliable (see section 2.4.3).

A machine learning (ML) framework is applied for the evaluation process of VAI. For the robustness of the inference, a feature vector needs to be formed by the features that have good discriminability between visual attention types in no-noise and acoustically noisy conditions. The features are compared in section 6.3.5, and the VAI performance is reported in section 6.5, which compares the inferring results with the labelled gaze data (i.e., supervised learning).

6.3 Feature Selection for VAI

6.3.1 Data labelling

In this evaluation, the task assumed by the system (TOVA) is to determine whether the user's fixation focus is the visual object he/she is talking about (i.e., looking at spoken objects). Reactive visual attention (RVA) is expected to be inferred by gaze events if the display changes with the system response and if the user moves eyes towards the change accordingly. For the other times, TIVA is supposed. In a 'WoZ' setting, the system response is controlled by the wizard.

The exaction of the gaze events is based on the sequences containing consecutive fixations and saccades. This event-based window approach is chosen against the time-based window approach. It is due to three main reasons. First, it is because the nature of the information events in gaze (e.g., fixations/VAs, saccades, pupil metrics, and so on) and speech (e.g., words, utterance, and so on): they do not occur in a fixed time interval or last for a fixed time duration. Compared to the time-based approach, this reduces the probability of separating a single event into two windows. Second, as fixation duration is an important variable in the feature vector, it is unwise to use a fixed time window here. Third, in regard to the real-time processing in the future, the event-based approach is also computationally effective since there will be less input data concurrently [21]. Figure 6.1 shows a typical example of the labelling process.

6.3.2 Feature extraction

The choice of gaze features normally leans to exploratory due to the lack of standardization [72] [137]. Three sets of measurements commonly used [263] are proposed to build the feature vectors (see section 2.2.2). The first set is the fixation focus and durations and the second set is the saccade measurements as suggested by a gaze feature review of Jakob [137]. The third set is the pupillary responses related to the cognitive load and

Sequence feature	Description
Event fixation duration	The duration of the fixation for on-going sequence
Prior fixation duration	The duration of the fixation before the event
Post fixation duration	The duration of the fixation after the event
Prior saccade length	The length of the saccade movement to the event fixation
Post saccade length	The length of the saccade movement from the event fixation
Average pupil size change	The average percent of change for the pupil size within the sequence

Table 6.1: Features computed from fixation, saccade and pupil size

photo sensitivity as stated in a study of Klingner [164]. For a gaze event, the prior and post events are also considered to form a richer feature set, resulting in a gaze sequence for feature extraction (Figure 6.1). These sets of features are listed in Table 6.1. The robustness of features are compared in section 6.3.5 in terms of their discriminability of different VA types in no-noise or acoustically noisy condition.

6.3.3 Feature normalisation

Fixations, saccade movements and pupil sizes, when used to build feature vectors in multi-users tasks, face limitations in utility due to the statistical distribution and the between people variation.

It is noticed that the gaze features are not normally distributed well (see section 5.5). This is due to the natural process limit, which is 0 in this case; because no fixation duration can be shorter than zero, a right skew can be observed from the distribution (Figure 6.3). As discussed in section 6.2, the process of VAI involves the prediction of $P(g_t = g|r_t = r)$. This density may be best estimated parametrically by a few parameters (e.g. a normal distribution with two parameters: mean and variance). Thus transforming the features to a normal distribution may assist robust prediction. As a standard pre-process step, the normalisation can potentially improve the machine learning result [122].

The variability between persons is caused by the variation of the users' biological and

emotional characteristics and the environmental illumination difference of the tasks [19]. Thus, a normalisation scheme over users is reasonable and necessary for the data to be comparable and aggregatable. The normalisation is important for the features to have equal weight during the classification process and to allow parametric descriptions.

To make the gaze data more normally distributed, a Box-Cox transformation [278] with $\lambda = 0$, which is a natural log transform, is applied on the features [4]. Figure 6.3 shows the distribution before and after the transform. It can be seen that the data are much better symmetrically distributed. Two methods are used to normalise features over the users. The first method is Z-score; the normalised value is retrieved by baseline subtraction and dividing the standard deviation [143]. As the samples in the data may not be sufficient enough to stand for true 'population', the z-score can also be treated as a student t-statistic here. The second method is the percent change, after the baseline subtraction, the partial result is divided by the baseline [17]. The first method is applied on fixation and saccade measures and the second method on pupil size. The overlaid histograms in Figure 6.4 show an example effect of the normalisation on event fixation duration.

6.3.4 Feature discriminability in no-noise condition

To assess the feature discriminability of VA types in no-noise condition, the estimation of parametric densities using normalised gaze features for the three defined VA types are listed in Figure 6.5. The results reveal the difference in fixation duration is higher when the user is instructing the system (TOVA mean 0.473 ms z-score) compared to when not (TIVA mean -0.593 ms z-score $p < 0.001$) or when reacting to changes in the visual field (RVA mean 1.530 ms z-score $p < 0.001$). Likewise, the prior saccade length shows significant differences ($p < 0.001$). The discriminability indicated by the significant difference makes them desirable features in the classification process.

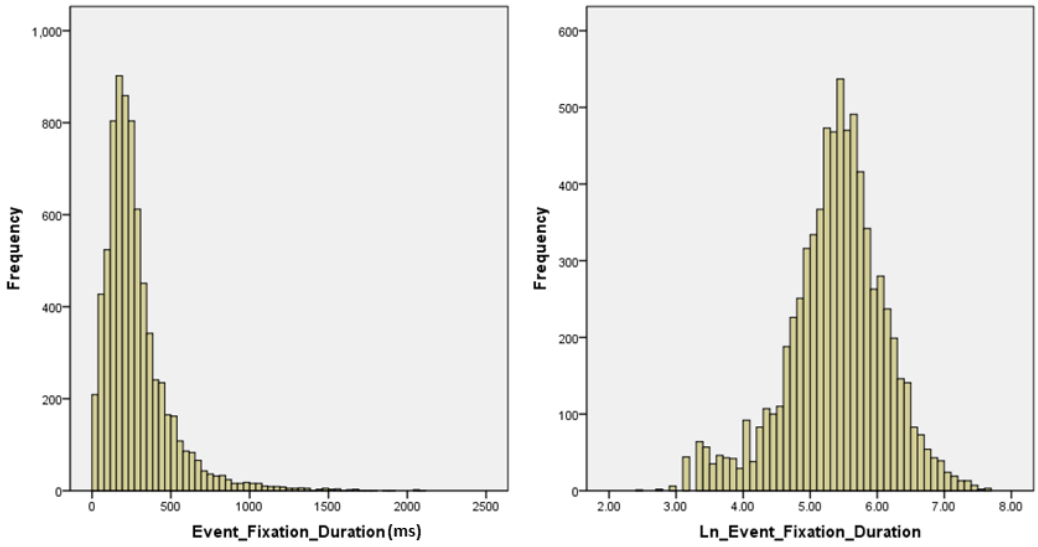


Figure 6.3: The overall fixation duration distribution before and after the nature log transform. The figure illustrates that after the transform, the fixation durations are much better symmetrically bell shaped.

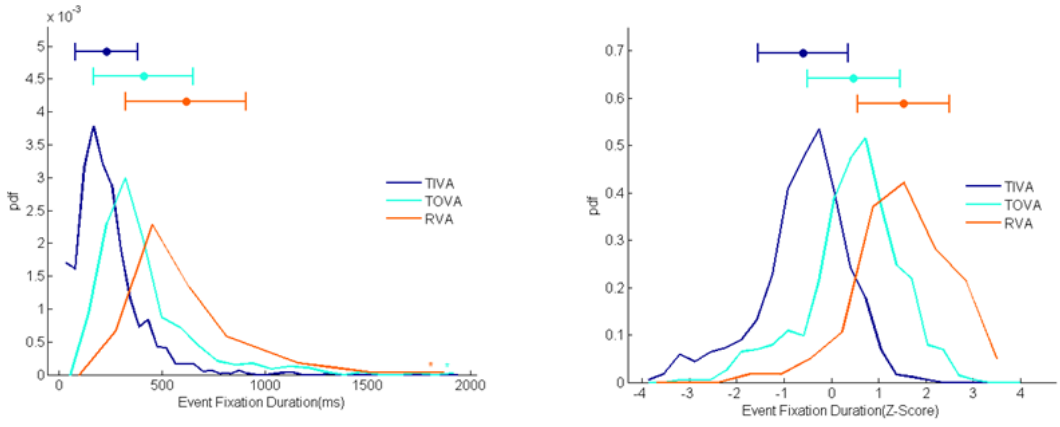


Figure 6.4: Difference in fixation duration distribution for each role before and after normalisation. A more symmetrical distribution results, which lends itself to parametric definition with less overlap in error bars between roles.

		Role		
		TIVA	TOVA	RVA
Event_Fixation_Duration (Z-Score)	Mean	-.593	.473	1.530
	Standard Deviation	.951	.976	.969
Prior_Fixation_Duration (Z-Score)	Mean	-.485	.095	1.039
	Standard Deviation	1.020	.925	.993
Post_Fixation_Duration (Z-Score)	Mean	-.130	-.096	.235
	Standard Deviation	.978	1.062	1.016
Prior_Saccade_Length (Z-Score)	Mean	-.236	1.003	-.426
	Standard Deviation	.994	1.101	1.006
Post_Saccade_Length (Z-Score)	Mean	-.004	-.046	-.008
	Standard Deviation	1.030	1.103	.961
Mean_Pupil_Size (percentage change)	Mean	.134	.117	.116
	Standard Deviation	.108	.095	.097

Figure 6.5: Summary statistics for normalised gaze features on a per-role basis without environmental acoustic noise. Fixation duration and its prior saccade length show the largest difference between roles, suggesting superiority over the other measures.

6.3.5 Feature discriminability and dependency upon acoustic noise

It is discussed in section 6.2 that the VAI is expected to be more beneficial for a gaze-contingent ASR in the acoustic noise. Thus, the discriminability of the features in noisy environment needs to be analysed.

In noisy environment, the 'gaze Lombard effect' is observed in the corpus data (see section 5.5). The general expectation is that the overall fixation duration will increase and that the overall saccade length will decrease during the noisy environment due to the psychological impact and the communication difficulty brought by the environmental noises [263]. In this section, the normalised features in different noise conditions regarding each VA type are investigated in terms of their discriminability of the VA types.

The gaze data is collected from the experiments in no-noise (N0), 42.75dB (N1), 54.75dB (N2), and 63.75dB (N3) noise environments respectively as discussed in the section 4.5.1. Two-tailed t-test is used to produce the significance value.

Event fixation duration (EFD)

As the noise increased, the prolonged event fixation duration (EFD) across all three roles can be noticed in Figure 6.6. The overall fixation duration increase in noisy environment is discussed in section 5.5. The results here reveal that this change is mainly caused by the increase of EFD within TOVA ($p < 0.001$) and RVA ($p < 0.001$), while the TIVA EFD does not significantly change ($p > 0.05$) in a noisy environment.

One plausible explanation for the increase of TOVA EFD is related to more active use of gaze to assist communication in noise. The observed increase of RVA EFD could be because that the user spends longer on evaluating the feedback of his commands in noisy environment.

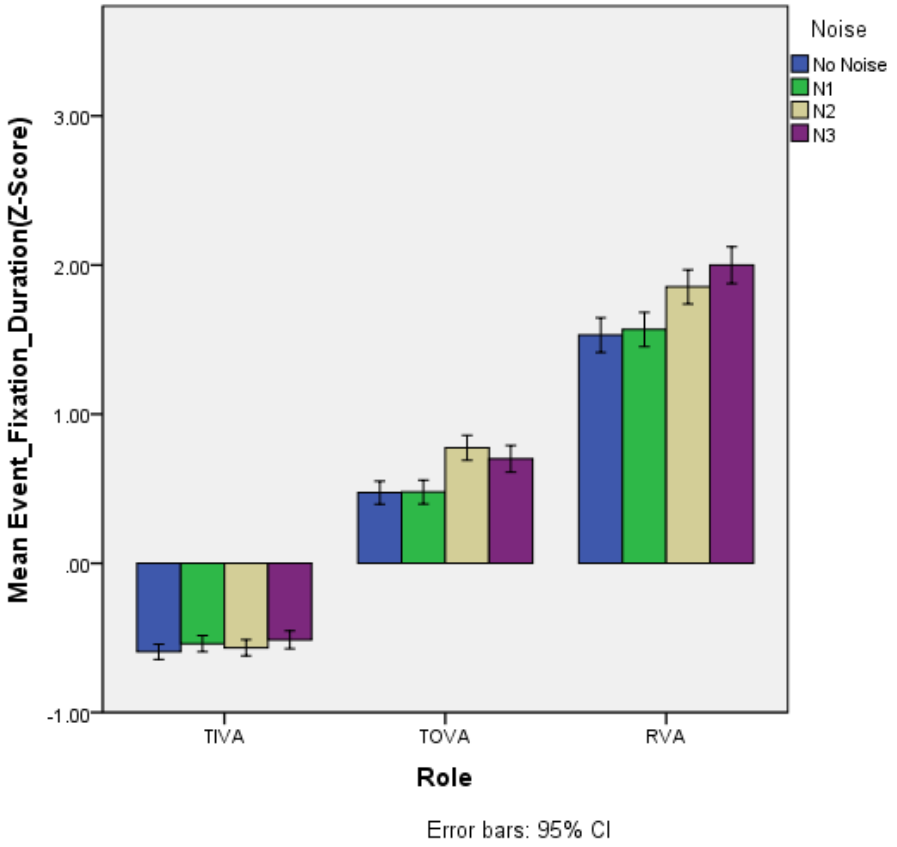
For the discriminability between VA types within each noise condition, the significant difference ($p < 0.001$) of EFD is observed between any two roles. The difference of distributions leads to the great potential for EFD to be a desirable feature in the classification process.

The experiment is repeated for prior and post fixation events because they may assist the classification process.

Prior fixation duration (PRFD)

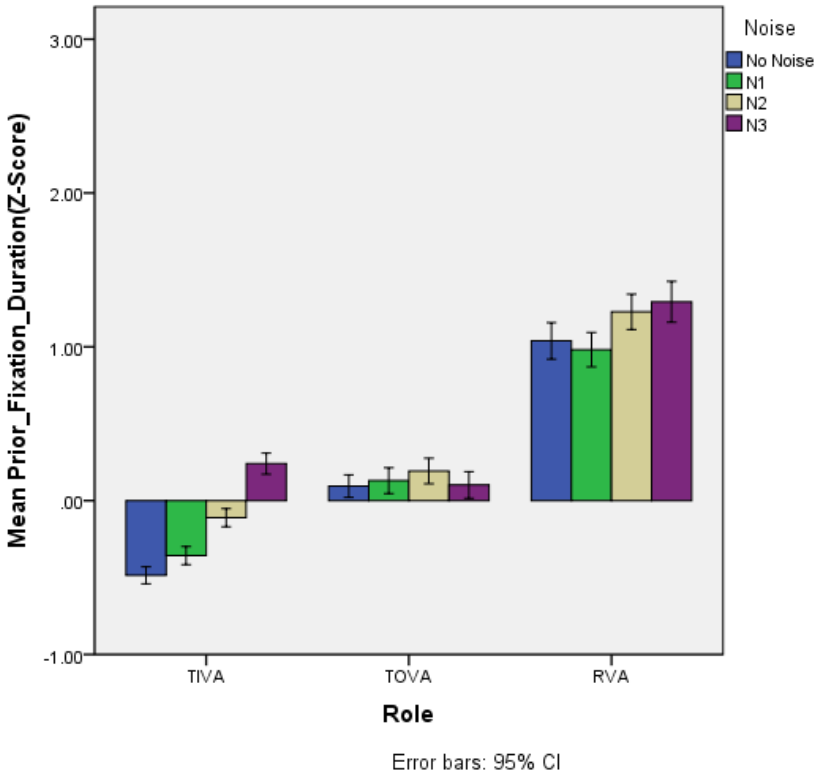
From Figure 6.7, an increase of the TIVA PRFD ($p < 0.001$) is noticed as noise increased, while the TOVA PRFD does not share this increasing trend ($p > 0.9$). A likely interpretation is that the PRFD of a TOVA sequence is related to when a user is planning for the next command to issue, and noise does not increase this duration where gaze is used for the psychological planning.

Within each noise condition, the RVA PRFD is longer than TIVA PRFD ($p < 0.001$) and TOVA PRFD ($p < 0.001$). A plausible explanation is that after issuing a command, a user is likely to retain his fixation focus waiting for the machine feedback. As the result of the distinguished distribution in each noise condition, PRFD has the potential to be a valuable feature in the classification process.



		Event_Fixation_Duration(Z-Score)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	-.593	.951	.473	.976	1.530	.969
	N1	-.539	.889	.478	.926	1.568	.964
	N2	-.566	.841	.775	.979	1.854	.943
	N3	-.513	.873	.700	1.154	1.999	1.020

Figure 6.6: The bar graph of the event fixation duration z-score with the 95% confidence interval error bar across all sequence roles and noise conditions.



		Prior_Fixation_Duration(Z-Score)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	-.485	1.020	.095	.925	1.039	.993
	N1	-.357	.956	.130	.967	.981	.934
	N2	-.111	.913	.192	.968	1.227	.943
	N3	.241	.992	.102	1.114	1.292	1.100

Figure 6.7: The bar graph of the prior fixation duration z-score with the 95% confidence interval error bar across all sequence roles and noise conditions.

Post fixation duration (PTFD)

In the Figure 6.8, no significant difference of PTFD is found between TIVA and TOVA either in no-noise condition ($p > 0.5$) or the noisiest condition ($p > 0.4$). Although the RVA PTFD is distinguished from TIVA ($p < 0.001$) and TOVA ($p < 0.001$) in no-noise condition, the difference becomes less notable in noise ($p > 0.4$). The unsatisfactory discriminability makes PTFD a less valuable feature in the classification process.

Prior saccade length (PRSL)

In section 2.2.2, the correlations between TIVA fixation duration and its prior saccade length are discussed. In Figure 6.9, this correlation can be identified from the TIVA PRSL. For the TOVA PRSL, a significant decreasing trend ($p < 0.05$) is seen. A likely interpretation is that while in noise, the disturbing environment makes a user tend to pick an object on the screen that is closer to the previous visual focus.

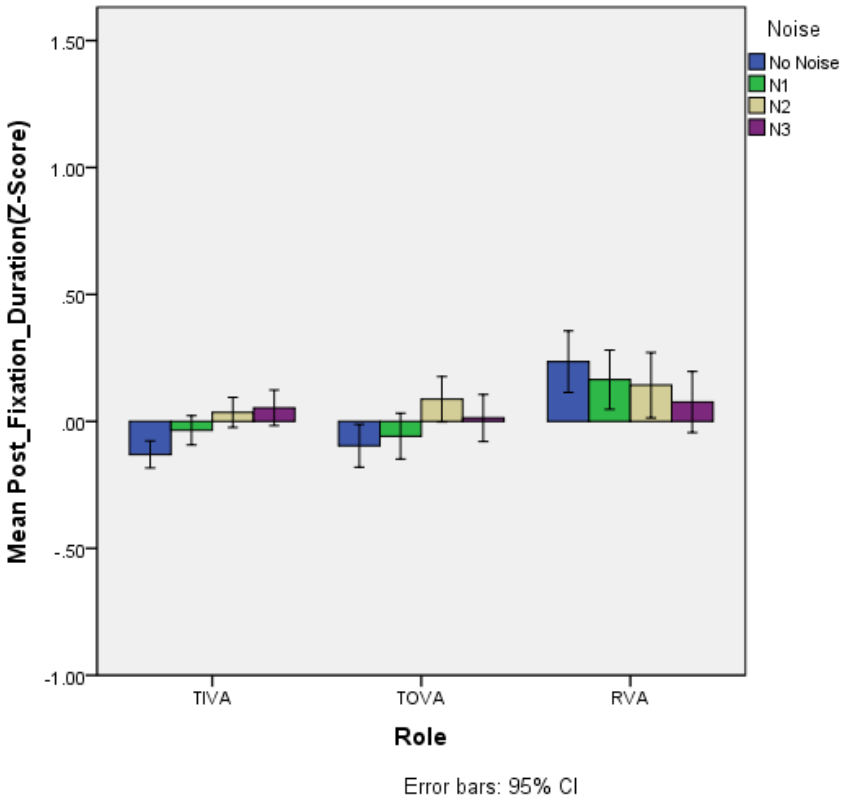
Within each noise condition, significant distinction ($p < 0.001$) exists between any two roles for PRSL, which makes it a good describing feature in the classification process.

Post saccade length (PRSL)

For the post saccade length (PTSL), the results in Figure 6.10 show that there is no significant difference ($p > 0.5$) between the distributions of any two roles in no-noise condition and between TIVA and RVA in noisy conditions ($p > 0.2$). In terms of discriminability, PTSL is less favourable than PRSL.

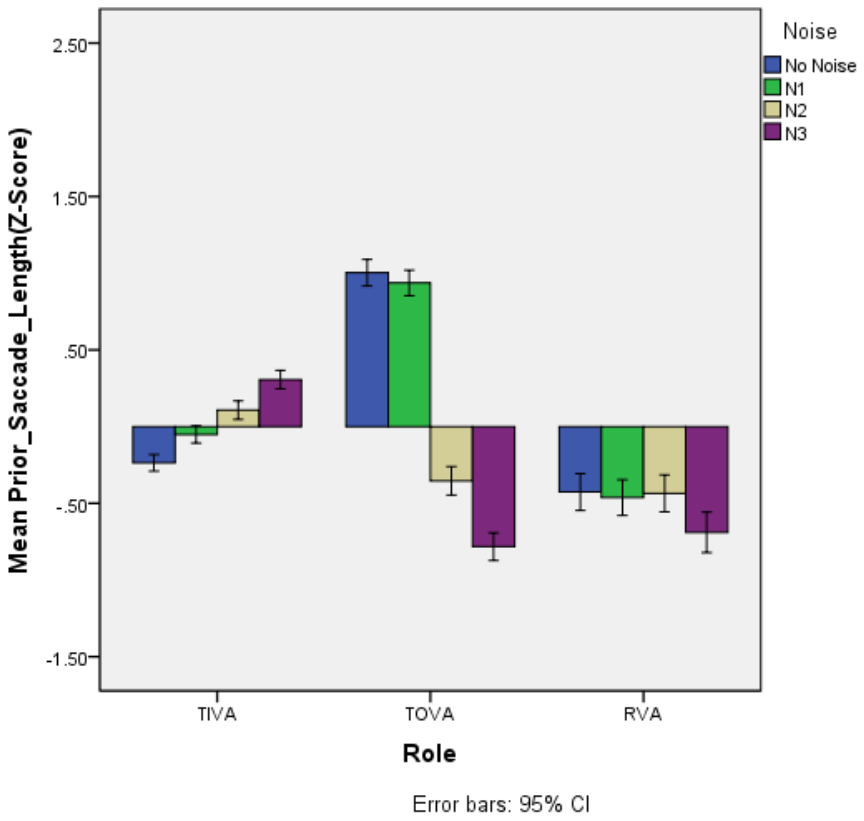
Pupil size change

The pupil size changes (%) are shown in Figure 6.11. From the results, an upward tendency can be observed for TOVA and RVA when noise increases ($p < 0.001$ between no noise and the noisiest condition). However, that there is no significant difference between VA types ($p > 0.05$) within noise conditions makes it less valuable in the VAI framework.



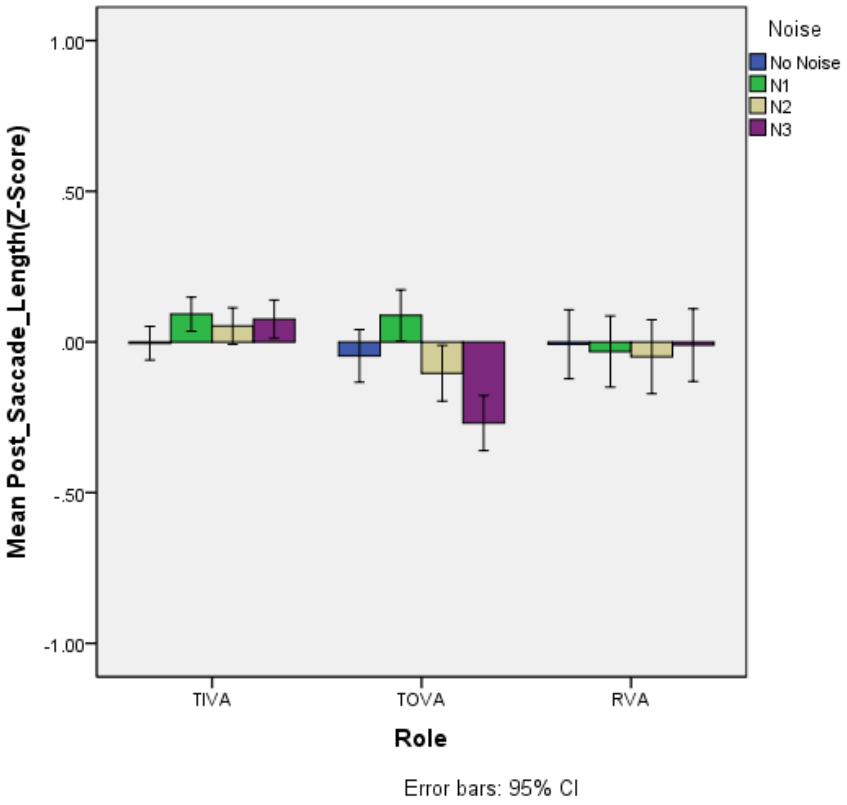
		Post_Fixation_Duration(Z-Score)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	-.130	.978	-.096	1.062	.235	1.016
	N1	-.035	.940	-.058	1.044	.164	.978
	N2	.036	.917	.087	1.044	.143	1.060
	N3	.054	1.017	.014	1.198	.076	.993

Figure 6.8: The bar graph of the post fixation duration z-score with the 95% confidence interval error bar across all sequence roles and noise conditions.



		Prior_Saccade_Length(Z-Score)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	-.236	.994	1.003	1.101	-.426	1.006
	N1	-.052	.917	.936	.968	-.462	.980
	N2	.107	.939	-.353	1.080	-.435	.990
	N3	.306	.874	-.783	1.159	-.689	1.097

Figure 6.9: The bar graph of the prior saccade length z-score with the 95% confidence interval error bar across all sequence roles and noise conditions.



		Post_Saccade_Length(Z-Score)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	-.004	1.030	-.046	1.103	-.008	.961
	N1	.092	.921	.088	.984	-.032	.994
	N2	.053	.935	-.104	1.083	-.049	1.010
	N3	.076	.918	-.269	1.178	-.011	.997

Figure 6.10: The bar graph of the post saccade length z-score with the 95% confidence interval error bar across all sequence roles and noise conditions.

This observation supports the previous finding that the size of pupillary change caused by external reflexes is distinctly larger than the change related to internal states [18].

Conclusion of findings

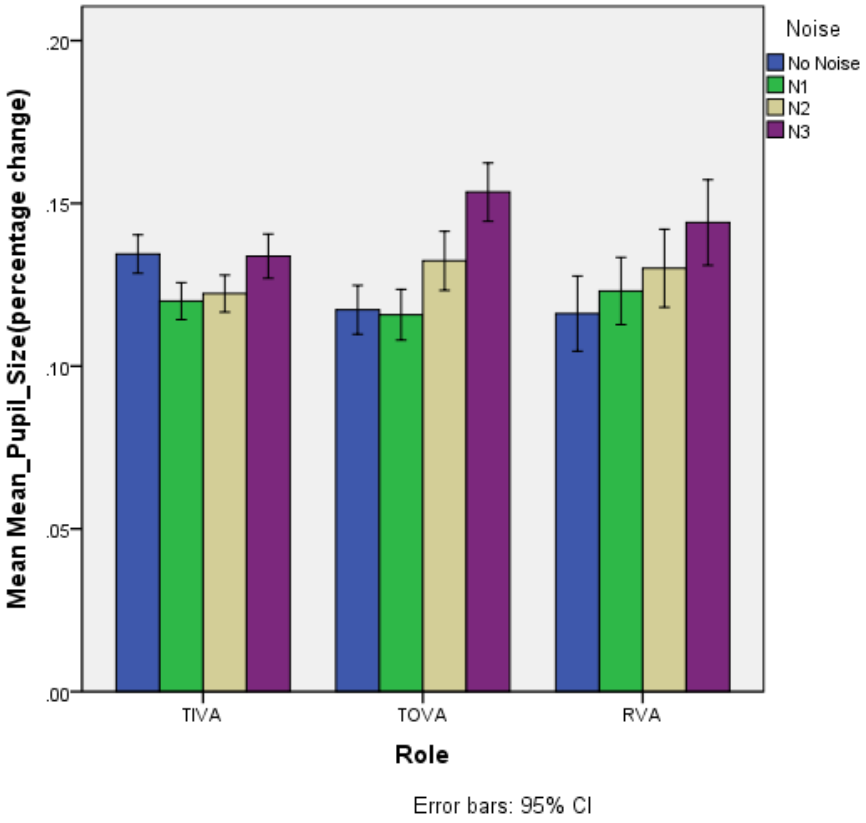
Based on the discussions above, event fixation duration (EFD), prior fixation duration (PRFD) and prior saccade length (PRSL) are more desirable features in the classification process compared to post fixation duration (PTFD), post saccade length (PTSL) and pupil size change. Thus, those features are used to form the feature vector for the classifier.

6.4 VAI Performance Test

To evaluate the VAI framework and the working taxonomy, a 3-class ML framework is applied to estimate $P(g_t = g|r_t = r)$ and $P(r_t = r)$ for predicting $P(r_t = r|g_t = g)$. The inferred results are evaluated using the classification performance metrics. To demonstrate the VAI framework in no-noise and acoustically noisy condition, the classification results in noise condition N0 and N3 are reported.

The features extracted in section 6.3.5 form a feature vector for the machine learning system. To derive the posterior probability density $P(r_t = r|g_t = g, M)$ (refer to expression 6.1), a naive Bayesian classifier is built as a prediction model. The estimated posterior probabilities $P(r_t = TIVA|g_t = g, M)$, $P(r_t = TOVA|g_t = g, M)$, and $P(r_t = RVA|g_t = g, M)$ are compared for the gaze event g_t with the feature vector to be inferred as one of the TIVA, TOVA, or RVA roles.

Table 6.2 shows the counts percentage of each participant and gaze role. It is noticed that the dataset is unbalanced as the TIVA data is approximately two times that of the TOVA data and four times that of the RVA data. This distinct class skew makes classification accuracy a poor performance metric choice for the evaluation. To better present the performance, area under receiver operating characteristics (AUC) is used



		Mean_Pupil_Size(percentage change)					
		Role					
		TIVA		TOVA		RVA	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Noise	No Noise	.134	.108	.117	.095	.116	.097
	N1	.120	.092	.116	.090	.123	.087
	N2	.122	.088	.132	.106	.130	.098
	N3	.134	.098	.153	.115	.144	.109

Figure 6.11: The bar graph of the average pupil size change with the 95% confidence interval error bar across all sequence roles and noise conditions.

		Role			
		TIVA	TOVA	RVA	Total
Participant	A	49%	33%	18%	291
	B	63%	30%	8%	342
	C	64%	24%	12%	217
	D	65%	22%	13%	315
	E	65%	23%	11%	374
	F	58%	30%	12%	356
	G	51%	36%	13%	291
Total		1297	617	272	2186
Total(%)		59%	28%	13%	100%

Table 6.2: Input data frequency. It is noticed that the dataset is distinctly unbalanced across VA types.

as a measurement insensitive to the class skew [82] in addition to other conventional measurements, such as precision, recall, and F-measure [252]. In this 3-class evaluation, AUC for each class is calculated in turn as a two-class situation by considering the other two as the negative class [254] for simple generation and visualisation. The averaged AUC reported is the weighted average of each class.

The performance of inferring gaze roles from the gaze characteristics alone (refer to expression 3.5) is compared with adding the multimodal coupling functions. In section 6.1, the multimodal coupling function between gaze and other modalities in the ES-N task are discussed. The temporal component of the coupling function is written as a decay function $f^t(.) = e^{-\Lambda|t-v|}$. In a decay function, the $\tau = 1/\Lambda$ is defined to be the lifetime (also called the exponential time constant). Of the interest for the study, the values of lifetime range from $0.1s$ to $8s$ (equally Λ ranges from 0.125 to 10) are investigated and the optimum τ is picked empirically based on the weighted average AUC. It is hypothesised that VAI is more reliable considering the multimodal coupling.

6.5 Results

To evaluate the system, 7-fold cross-validation is employed with each fold contains the data from a participant. Using the gaze characters alone, the classification results in

	Classification Results	
Features	Accuracy	Avg AUC
Pupil Size	0.593	0.532
Saccade Length	0.711	0.720
Fixation Duration	0.735	0.832
S+P	0.709	0.719
F+P	0.734	0.829
F+S	0.808	0.890
F+S+P	0.809	0.890

Table 6.3: Classifier performance for gaze role inference in no noise environment demonstrating the benefit of using fixation duration (F) and saccade length (S) over pupillary responses (P)

Table 6.3 confirm that for no-noise condition, fixation duration and the length of its prior saccade has 81% accuracy, demonstrating excellent performance in terms of the AUC (0.89). Similarly, for noisy condition, the best accuracy and AUC are 75.3% and 0.892 respectively with fixation duration and saccade length as part of input features.

To incorporate the multimodal coupling function, the optimum decay function parameter Λ is chosen empirically from values ranged from 0.125 to 10 to ensure the optimum average AUC. During this process, the word sequence obtained by ASR and the recorded screen-display-change events were used. The results are shown in Figure 6.12 for classification in two noise conditions, clean and noisy. The $\Lambda = 1$(lifetime $\tau = 1s$) is chosen for the best classification performance.

The detailed improved performances after incorporating the coupling function for both conditions are listed in Table 6.4. While the other evaluation measures are more sensitive to the class skew (unbalanced class distribution), AUC is more of the interest in this case. After incorporating the coupling function, the TOVA AUC increases from 0.847 to 0.866 and from 0.839 to 0.844 for no-noise and noisy conditions respectively. Similarly, the RVA AUC increases from 0.946 to 0.970 and from 0.923 to 0.946 respectively. Meanwhile, there is no change for the TIVA AUC. Together, they lead to an improvement for the weighted averaged AUC from 0.890 to 0.903 in no-noise condition and from 0.892 to 0.898 in noisy condition.

It is observed that incorporating the coupling function assists to resolve the ambiguity

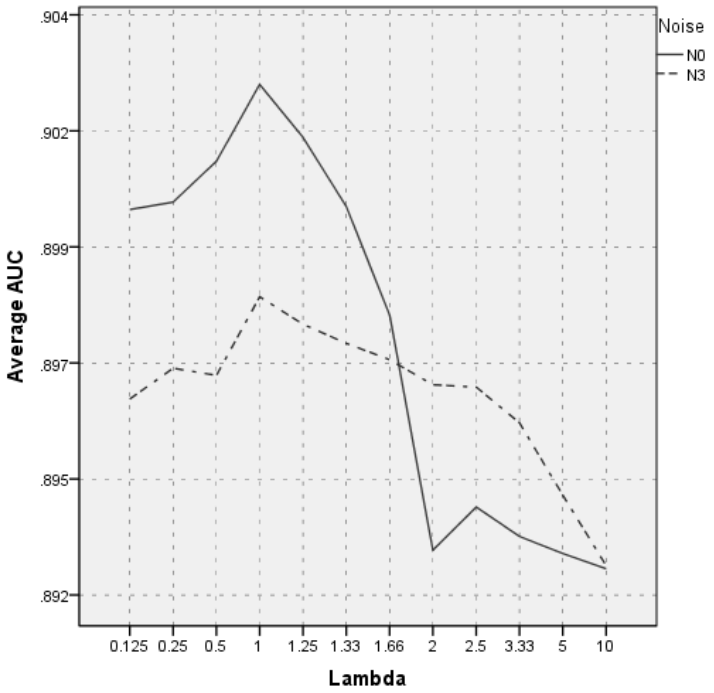


Figure 6.12: The weighted average AUC of the classifier for no-noise and noisy conditions when decay rate Λ in the coupling function ranges from 0.125 to 10.

between the classification for TOVA and RVA gaze events. This can be revealed in the comparison between the confusion matrix in Table 6.5 for no-noise condition and Table 6.6 for noisy condition. Table 6.5 shows that the cases in which TOVA is misclassified as RVA drop by 8.3% (from 48 to 44) and the cases in which RVA is misclassified as TOVA drop by 9% (from 44 to 40). The corresponding figures in noise condition are 19.7% and 5.3% respectively, as shown in Table 6.6.

6.6 Summary

In Chapter 5, the semantic and temporal relationship between speech and gaze is investigated for the inference of acoustic noise, and the demand to adapt an event-based inferring framework in ASR systems for appropriate integration of gaze is raised. In this chapter, a visual attention inference (VAI) framework is described as an event-based inferring function to account for the relevance of gaze events to system function.

Noise Condition	N0		N3	
Incorporate f	No	Yes	No	Yes
Accuracy	0.809	0.812	0.753	0.769
Precision	0.806	0.819	0.750	0.767
Recall	0.809	0.822	0.753	0.769
F-Measure	0.807	0.820	0.750	0.767
Avg AUC	0.890	0.903	0.892	0.898
TIVA AUC	0.894	0.896	0.924	0.923
TOVA AUC	0.847	0.866	0.839	0.844
RVA AUC	0.946	0.970	0.923	0.946

Table 6.4: The detailed performances before and after incorporating the coupling function f. The latter demonstrates an overall improved performance.

	Predicted Class		
	TIVA	TOVA	RVA
TIVA	1151	108	38
TOVA	160	409	48
RVA	20	44	208
TIVA	1159	102	36
TOVA	167	406	44
RVA	22	40	210

Table 6.5: The confusion matrix for no-noise condition before (top) and after (bottom) incorporating the coupling function.

	Predicted Class		
	TIVA	TOVA	RVA
TIVA	698	105	2
TOVA	157	418	66
RVA	17	75	173
TIVA	697	100	8
TOVA	151	437	53
RVA	12	71	182

Table 6.6: The confusion matrix for noisy condition before (top) and after (bottom) incorporating the coupling function

Within the VAI framework, a gaze role taxonomy is proposed aiming at distinguishing task-relevant gaze events with others. The formalism of the VAI and coupling function is discussed and applied to the ES-N corpus task.

As the task assumed by the system (TOVA) is determined by the system function, the evaluation is conducted, aiming at more appropriate integration with speech using the relevant information in gaze (e.g., to improve speech recognition in noise). The evaluation using the gaze-speech corpus data collected in Chapter 4 shows support for including a VAI function in interactive systems. A naive Bayes classifier is shown to perform well in either no-noise condition or noisy condition with the gaze characteristics alone.

Related to the need for richer feature sets, the VAI framework proposed in this chapter incorporates multimodal coupling functions. With the addition of coupling functions, the performance improvement is illustrated within all the evaluation metrics (accuracy, precision, recall, f-Measure, and AUC) in both noise conditions stating the value of incorporating coupling functions in VAI. VAI will be used for the selective use of VA information in language model adaptation in the ASR system.

To validate the value of implementing VAI and ANI framework in engineering an ASR system, an application example is built in Chapter 7 to demonstrate the performance improvement.

CHAPTER 7

SELECTIVE GAZE-CONTINGENT ASR SYSTEM

In Chapter 5, an acoustic noise inference (ANI) framework is described by exploring the relationship between gaze and speech and the dependency upon noise. In Chapter 6, a visual attention inference (VAI) framework is proposed for highlighting the benefit of distinguishing gaze events by the visual attention type. The evaluation is conducted assuming a system function of integrating gaze with speech for a better recognition in acoustically noisy environment.

In this penultimate chapter, an ASR system is constructed. The acoustic model and language model adaptation techniques are used to improve the ASR performance. The former is performed based on the noise condition inferred, and the latter is performed based on the visual attention type inferred. For this purpose, the implementation of ANI and VAI frameworks in this ASR system is described. Because the ASR uses gaze information selectively, it is termed to as a selective gaze-contingent ASR.

7.1 ASR and Adaptation Overview

7.1.1 ASR basics

Automatic speech recognition (ASR) can be defined as a technology that allows the computer to identify a user's speech and transcribe it into readable text (see section 2.3.1).

A popular ASR system consists of two components: an acoustic component known as acoustic model and a linguistic component that incorporates a language model [170] (See Figure 7.1). An acoustic model takes acoustic inputs from the speech and compiles them into statistical representations of the sounds. It then matches them with the words in the vocabulary and assigns a probability to each word. One most common acoustic model is the hidden Markov model (HMM) [260]. A language model estimates the probability of a word occurring based on the history of previous words. The overall probability of a word candidate in ASR is then calculated based on the combination of the two probabilities produced by acoustic model and language model respectively.

More formally, let A denote the acoustic evidence observed; the objective of ASR is to find the most likely word sequence $\hat{W}$ by Bayesian inference:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|A) = \underset{W}{\operatorname{argmax}}\, P(A|W)P(W) \tag{7.1}$$

where $P(W|A)$ denotes the probability of word sequence W is spoken given the evidence A is observed. While the conditional probability of the acoustic evidence being observed $P(A|W)$ is calculated by the acoustic model, the prior probability of word sequence $P(W)$ is provided by the language model.

7.1.2 Language Model

A language model assigns a probability to a sequence of N words $W = (w_1, w_2, ..., w_N)$. The probability of W can be expressed as a product of conditional probabilities:

$$P(W) = \prod_{i=1}^{N} P(w_i|w_{i-n+1}^{i-1}) \tag{7.2}$$

Within the term $P(w_i|w_{i-n+1}^{i-1})$, $w_{i-n+1}^{i-1} = (w_{i-n+1}, \ldots, w_{i-2}, w_{i-1})$ is considered the history of n previous words and w_i the prediction. In an n-gram model, two histories are considered identical if they end in the same $n-1$ words. The histories are estimated from the speech data. The conditional probability $P(w_i|w_{i-n+1}^{i-1})$ in an n-gram language

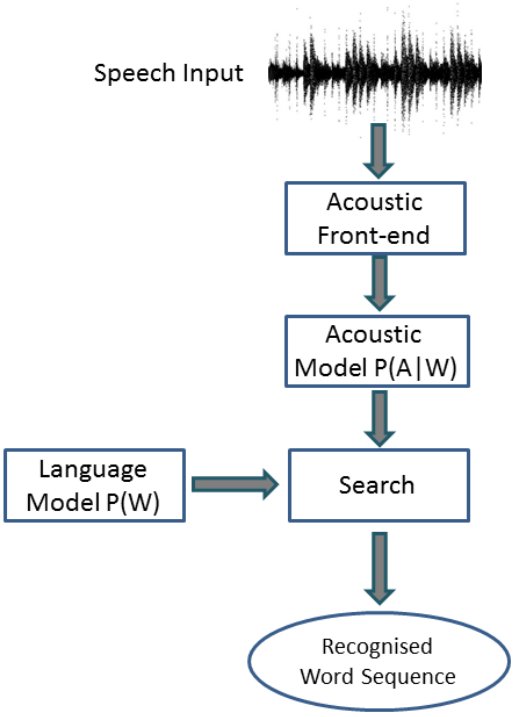


Figure 7.1: A standard basic model of automatic speech recognition that involves the use of an acoustic model and a language model.

model can be estimated with the maximum likelihood estimation (MLE) technique where C denotes the frequency count:

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})} \tag{7.3}$$

In bigrams, with $n = 2$, the probability of a word sequence W becomes:

$$P(W) = \prod_{i=1}^{N} P(w_i|w_{i-1}) \tag{7.4}$$

And the conditional probability $P(w_i|w_{i-1})$ is expressed as:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \tag{7.5}$$

7.1.3 Selective gaze-contingent ASR architecture

With the ANI and VAI frameworks evaluated in Chapter 5 and Chapter 6 respectively, the system described in Chapter 3 can be applied as the selective gaze-contingent ASR architecture shown in Figure 7.2.

With the acoustic noise inferred using ANI, various strategies can be used to counter the effect brought by the noise. These strategies include speech enhancement, feature enhancement, model adaptation, noise-resistant feature extraction (see section 2.3.4). In this study, for the demonstration of the value of ANI in ASR systems, an acoustic model adaptation technique is used.

The gaze events are used selectively in a cache-based LM adaptation based on the visual attention type inferred by the VAI framework. As discussed in section 2.4.3, it is expected that the use of gaze will be more effective and valuable in the noisier environment. It needs to be noted that the selective use of gaze can be applied to the ASR system regardless of the preceding counter-noise strategy used.

7.1.4 Gaze-based LM adaptation

Language model (LM) adaptation is the process of modifying the word probabilities in a LM trained in speech from one domain to better model speech in another domain (e.g., a topic). In a previous study by Cooke [49], LM adaptation based on the speaker focus of visual attention at time t modified the probabilities of words associated with map objects viewed. It was demonstrated that in the LM adaptation process, the redistribution of probability mass in a multiple-class LM yielded better WER performance improvement than a single class LM; one class contained words associated with the visual field, and probability mass was redistributed only within that class in response to a gaze event.

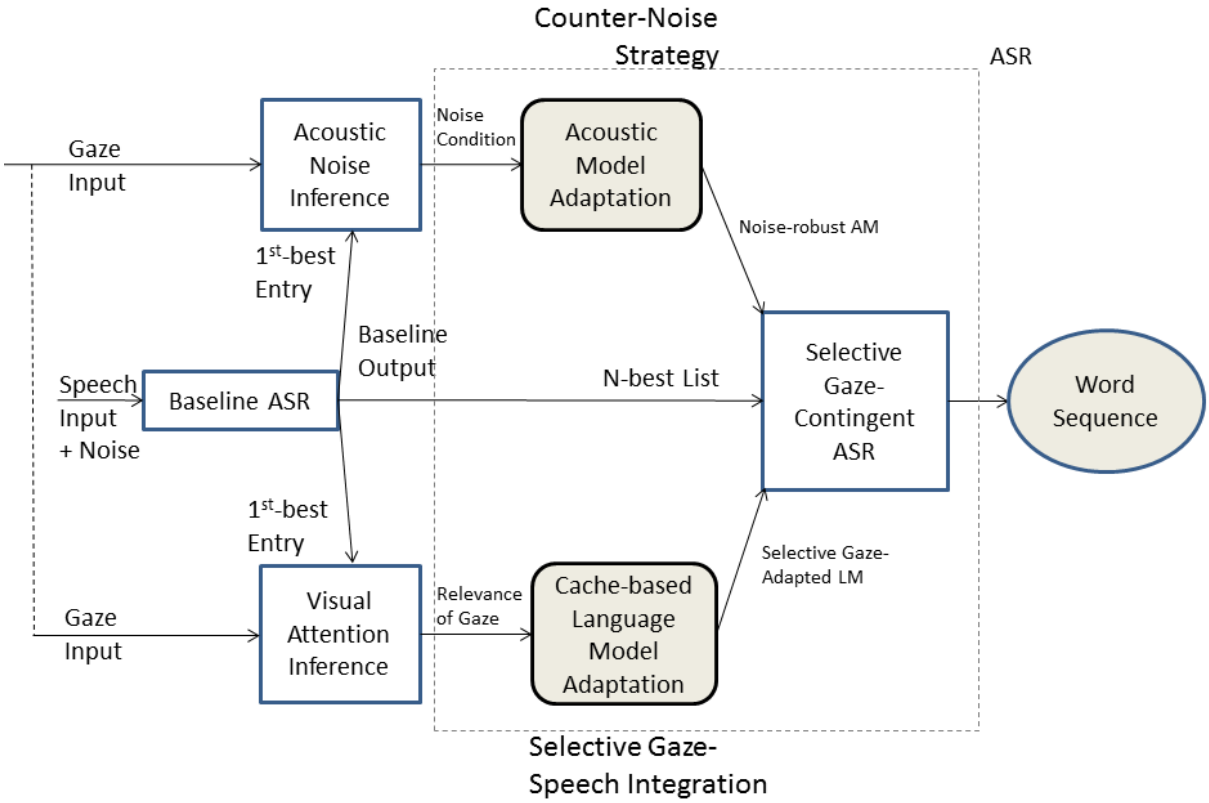


Figure 7.2: The implementation architecture of the selective gaze-contingent ASR. The gaze is used selectively based on the VAI results in cache-based LM adaptation for the integration with speech. The counter-noise strategy used in the study is the acoustic model adaptation for demonstrating the value of the ANI framework. The selective gaze-contingent ASR utilises these adaptation techniques to improve the noise-robust recognition performance.

7.1.5 Cache-based LM adaptation

Cache-based LM adaptation utilises a cache of previous events to boost the probabilities of words occurring in the cache. Typically, the cache at time t contains the previous hypothesised words up to time t. For example, Kuhn [170] proposed a cache model applied on a class-based LM [31] in which the word probability $P(w|C)$ of class C is updated by the recent 200 words. LM perplexity and WER improvements have also been shown by using topic-based cache models [45] [215] where previous hypothesised topics of conversation form the cache rather than words.

In addition to linguistic-derived caches, caches based on vision contain a reference to physical objects in the environment [183] or in the users' field of view [274]. The object references in these caches are associated with keywords, which are boosted in LM adaptation relative to their cache occurrence; Qu [257] employs this technique in his study and shows decreases in LM perplexity. In another LM adaptation study by Cooke [50], a cache is proposed containing gaze events (fixations on visual foci) before and after a hypothesised word based on assumptions of psycholinguistic processes. However, in these studies, WER improvements are limited, and the systems are not evaluated in noisy environments (see the discussion in section 2.4.3).

The technique described in this chapter builds on previous work by investigating the selective use of gaze in cache LM adaptation of a class-based LM, where the selection criteria for gaze events are learnt and implemented using the VAI framework discussed in Chapter 6. An N-class LM model extends the 2-class model proposed by Cooke [49] so that multiple classes are used to represent the visual field and task. The evaluation is conducted using eye movement and speech data recorded in acoustically noisy environment (ES-N corpus, see Chapter 4) to better match real-world utility.

7.1.6 Acoustic model adaptation

An acoustic model is trained using a particular set of speech data. When there is a mismatch between the training condition and recognition condition, an adaptation procedure can reduce the mismatch and improve the recognition performances (discussed in section 2.3.4).

For an HMM acoustic model, one typical adaptation approach is the re-estimation of the HMM parameters. Among the adaptation techniques, maximum likelihood linear regression (MLLR) [175] and maximum a posteriori (MAP) [97] are most popular. While MLLR helps in dealing with unseen models, MAP uses prior knowledge to account for speaker variation; therefore, it is particularly useful in dealing with informative prior knowledge (i.e., knowing what the parameters of the model are likely to be using the prior knowledge).

The MLLR computes a linear transformation $\hat{\mu}$ of the mean vectors μ of the Gaussian densities with the transformation matrix W and bias b

$$\hat{\mu} = W\mu + b \tag{7.6}$$

to optimise the maximum likelihood by maximising the auxiliary function [16] Q:

$$Q(\mu, \hat{\mu}) = \sum_{\theta \in S} F(O, \theta|\mu)\, log(F(O, \theta|\hat{\mu}) \tag{7.7}$$

where θ denotes the sequence of states to generate observation O, S the set of all possible state sequences, and F the function of likelihood.

For the MLLR, a global transform can be applied to every Gaussian component in the acoustic model, or more specific transforms can be applied to certain subsets of the model. In the latter situation, a *regression class tree* can be used to group the components that are close in acoustic space [325] and therefore be referred as 'regression MLLR'. MLLR is particularly useful in the situation where adaptation data is limited as it estimates a

transform matrix to apply on acoustic models instead of re-estimate the model parameters directly. It is believed to perform better accounting for different recording apparatus and environment [51] but has also been used for speaker adaptation [175].

On the other hand, MAP re-estimates the mean vector μ by maximising the posterior probability p given the observation O

$$\mu_M = \underset{\mu}{\operatorname{argmax}}\, p(\mu|O) \tag{7.8}$$

MAP performs better with large amounts of data and is good at dealing with speaker variation as opposed to recording apparatus and environment [51]. It is possible to take advantage of both techniques by serialising them [62] [49], i.e., MLLR followed by MAP.

7.2 Framework for Selective Gaze Integration

This section presents a general framework for selective use of gaze in cache adaptation of class-based N-gram LMs. A cache-based adaptation of a class-based language model is represented, with the cache containing gaze events instead of word or topic events, and classes formed by considering how gaze information relates to information in speech.

7.2.1 Baseline LM construction using class-based model

In a language model, words can be clustered together into an equivalence class and this would result in an n-gram class model [32]. For example, if the n-gram word probabilities of two persons' name, Tom and Jack, are comparably relative in the vocabulary, they can be clustered into a class (e.g., name class) and treated as equivalent for the language modelling. Standard word-based n-gram models can be considered a special case of class-based models in which every single word is mapped to a unique word class.

For the baseline language model (LM) a class-based n-gram model is used:

$$P_b(W) = \prod_{i=1}^{N} P_b(c_i|c_{i-1}^{i-n+1}) P_b(w_i|c_i) \tag{7.9}$$

where N is length of the word sequence $W = \{w_1, w_2, \ldots, w_N\}$, c_i the class, w_i the word, and suffix 'b' the 'baseline'. $c_{i-1}^{i-n+1} = \{c_{i-n+1}, \ldots, c_{i-2}, c_{i-1}\}$ is the history of n previous classes associated with the n previous words. The class-based model enables a visual task-specific grammar to be captured - e.g., words associated with groups of visual foci, where the grouping of foci is related to task. It also overcomes the sparseness problems by estimating infrequent words that have support from the frequent ones in the same class [215]. In a previous study [274], the class-based model is reported to require only about one-third as much storage as the standard language model, in which each word is treated as a unique individual.

7.2.2 Cache-based LM adaptation

The baseline LM word probabilities $P_b(w_i|c_i)$ are time invariant. Cache-based LM adaptation is used to modify the LM at time t given extra information in a cache. Figure 7.3 illustrates the basic idea. Assume there are in total M gaze events up to time t, $G_{total} = \{g_1, \ldots, g_M\}$. The cache is formed from the sequence of the latest l gaze events, $G_{cache} = \{g_{M-l+1}, \ldots, g_M\}$, where l is the cache length (i.e., history of gaze events). The gaze event cache LM word probabilities, $P_g^t(w_i|c_i)$ are computed at time t as:

$$P_g^t(w_i|c_i) = \frac{\sum_{m=M-l+1}^{M} \sigma(g_m, w_i)}{\sum_{m=M-l+1}^{M} \sigma(g_m, c_i)} \tag{7.10}$$

where $\sigma(g_m, w_i)$ is a function that represents the relevance of the gaze event g_m to the word w_i, and $\sigma(g_m, c_i)$ is a function that represents the relevance of the gaze event g_m to the class c_i. These *relevance functions*, $\sigma(.)$ determine whether and to what degree the gaze event modifies the word probabilities. Thus, the word sequence probability becomes:

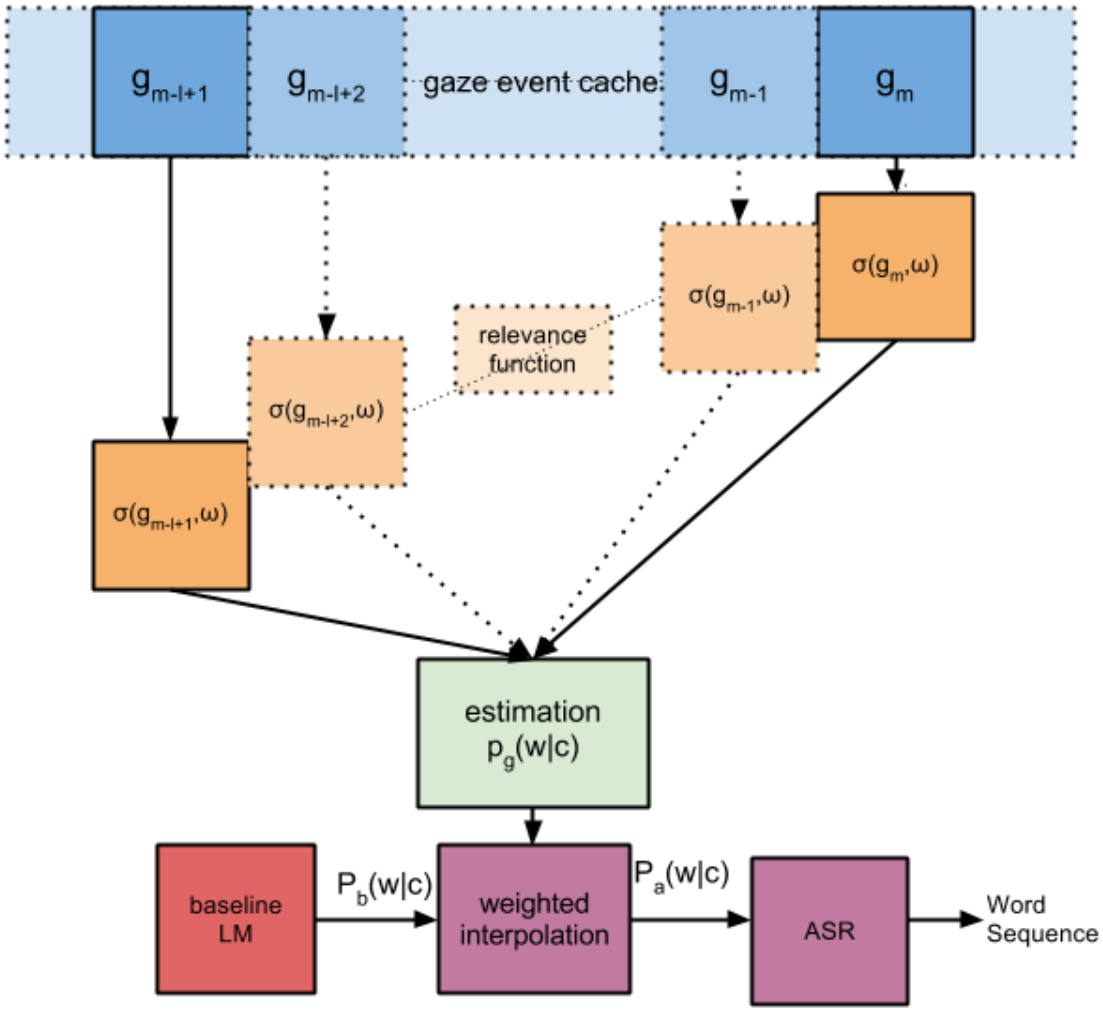


Figure 7.3: Cache-based LM adaptation framework. A cache is made of gaze events selectively (via relevance function) to adapt the class-based LM. The arrows represent the data flow.

$$P_g^t(W) = \prod_{i=1}^{N} P_b(c_i|c_{i-1}^{i-n+1}) P_g^t(w_i|c_i) \tag{7.11}$$

The adapted LM word probability at time t, $P_a^t(w_i|c_i)$, is determined from the weighted interpolation of the baseline LM $P_b(W)$ and the cache LM $P_g^t(W)$:

$$P_a^t(W) = (1-\lambda)P_b(W) + \lambda P_g^t(W) \tag{7.12}$$

The interpolation parameter λ given in expression 7.12 enables information from gaze to be used *in toto* to adapt the LM.

7.2.3 Relevance function

The relevance functions $\sigma(g, w)$ and $\sigma(g, c)$ represent the degree to which a specific gaze event g is related to the word w and class c respectively. For easier representation, $\sigma(g, \omega)$ is defined where ω is either class or word. When a TOVA is inferred by the VAI framework, a confidence score s will also be produced. For the non-selective use, a TOVA is treated equally as the other VAs. This approach is comparable with the ones used by Qu [257] and Cooke [50]. For the selective use, a TOVA is considered relevant and should be used in the cache with the degree of relevance weighted. However, in score-based selective use, the weight is the confidence score s; while in definitive selective use, all TOVAs are weighted equally (i.e., classification confidence is not considered). Consequently, three approaches of estimating the relevance function for integrating gaze with speech are proposed.

Relevance function 1: Non-selective use

In relevance function 1, all gaze events associated with ω are considered equally relevant in the cache. The function $\sigma(g, \omega)$ in expression 7.10 is defined as:

$$\sigma(g,\omega) = \begin{cases} 1 & \text{if } g \text{ is related to } \omega \\ 0 & \text{otherwise.} \end{cases} \tag{7.13}$$

Relevance function 2: Score-based selective use

In relevance function 2, the relevance of gaze events is variable and measured by a score s. The function $\sigma(g,\omega)$ in expression 7.10 is defined as:

$$\sigma(g,\omega) = \begin{cases} s & \text{if } g \text{ is related to } \omega \\ 0 & \text{otherwise.} \end{cases} \tag{7.14}$$

where s is a continuous value representing the measure of relevance between 0 (no confidence) and 1 (full confidence). In this work, the score s is calculated by the VAI framework (will be described in the next section).

Relevance function 3: Definitive selective use

In relevance function 3, the relevance of the gaze event depends upon a definitive decision i.e., a classifier decision. The function $\sigma(g,\omega)$ in expression 7.10 is defined as:

$$\sigma(g,\omega) = \begin{cases} 1 & s \geq \gamma \text{ and } g \text{ is related to } \omega \\ 0 & \text{otherwise.} \end{cases} \tag{7.15}$$

where γ represents a threshold for the classifier to classify the gaze event as a TOVA.

7.2.4 VAI implementation for relevance functions

In the section 7.2.3, three approaches have been proposed to account for the relevance between speech and gaze for integration. In *Approach 1: Non-selective use*, a gaze event g in the cache is used non-selectively to bias the LM to increase the probability of the

related word w. While in *Approach 2: Score-based selective use* and *Approach 3: Definitive selective use*, the score s needs to be calculated to represent the confidence of a gaze event g relevant to speech in the case of the system task being looking at the spoken words (i.e., the system task assumed in VAI evaluation section 6.2).

Referring back to section 6.1, the score s can be determined by the VAI result $p(r_t = TOVA|g_t = g, M) = p(r_t = TOVA|\overline{X}, M)$, where $\overline{X}$ denotes the inference feature set (see section 6.3.5) of g, and M denotes the other modalities (e.g., speech utterance and system response). Therefore, the score s is calculated by the VAI framework as:

$$\begin{aligned} s = &\ p(r_t = TOVA|\overline{X}, M) \\ \propto &\ p(\overline{X}|r_t = TOVA)\ p(r_t = TOVA)\ p(f_{TOVA}(M, g_t = g)|r_t = TOVA, g_t = g) \end{aligned} \tag{7.16}$$

The threshold γ in *Approach 3* is determined for the gaze event g to be classified as TOVA, i.e.,:

$$\gamma = \max(\ p(r_t = TIVA|\overline{X}, M),\ p(r_t = RVA|\overline{X}, M)\ ) \tag{7.17}$$

7.3 Evaluation Methodology

7.3.1 Method

A full ASR system is built that implements the frameworks, notably the VAI approach for implementing the relevance function. The performance of the ASR is evaluated using the ES-N corpus collected in different background noise conditions, as described in Chapter 4.

7.3.2 Baseline LM and class construction

The baseline LM is constructed from the speech transcriptions of the ES-N corpus data containing 1056 utterances and 3764 words with a vocabulary of size 91. To feed the

language model to the ASR in the later stage, bi-grams model is used. For the smoothing technique, Witten-Bell [43] is adopted for it has been reported to outperform Good-Turing, linear and absolute smoothing in bi-grams with small corpus data [95].

With the prior knowledge of the task involving user instructing positioning of coloured shapes (e.g., 'red square at left'), four classes are created in the LM - colour, shape, position, and non-visual-related. Words related to these classes are assigned to them. This implementation is used to reflect the task-specific relation between words and visual foci.

7.3.3 Training the baseline ASR

The construction of a proper ASR requires supervised training on a corpus. The ASR is trained on WSJCAM0 corpus of British English [271]. WSJCAM0 corpus was recorded at Cambridge University. Derived from the Wall Street Journal text corpus, it is of the largest corpora of spoken British English. The pronunciation dictionary with 25231 triphones in the ASR was derived from the British English Example Pronunciation (BEEP) dictionary, which is part of the WSJCAM0 corpus.

The ASR performance on the WSJCAM0 test data reached a 20.9% WER, which shows favourable performance against the systems in other studies [240] [130] [305]. The audio input was transformed using static and dynamic Mel frequency cepstral coefficient [61] with cepstral mean normalisation [10]. The ASR was built using the HTK [325] toolbox on a Linux system. More details of this baseline ASR may be found in a previous study by Cooke [49].

7.3.4 Parameter selection

In expression 7.1, the most likely word sequence $\hat{W}$ was expressed simply as the product of acoustic and linguistic probabilities. However, in real systems, more efforts need to be paid on the balance between acoustic and linguistic parameters to optimise the sys-

tem performance [299]. This is done by introducing a language model weight (so called language model scaling factor LMSF) and word insertion penalty (WIP):

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ P(A|W)P(W)^S\ i^N \tag{7.18}$$

where s denotes the LMSF, i the WIP, and N the number of words in sequence W. Normally, the computation is carried out in log domain:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ \log P(A|W) + s \log P(W) + N \log i \tag{7.19}$$

The LMSF affects the transition between words as it multiplies the transition probabilities by a constant. As LMSF is increased, there will be more deletion errors and fewer insertion errors. Also, a larger LMSF would mean less influence by acoustic model observation probabilities.

The WIP also functions as a penalty for inserting words as the name suggests. It controls the trade-off between insertion and deletion errors. A larger penalty makes the decoder prefer fewer longer words and, therefore, brings more deletion errors and less insertion errors. A smaller penalty makes the decoder prefer more shorter words and, therefore, has the opposite effect. It needs to be noticed that a larger penalty is introduced by a smaller (more negative) WIP value.

The LMSF and WIP are important towards an optimised ASR performance, and they are decided empirically using the grid-search. Figure 7.4 shows the ASR performance (WER) on an example speech recording data as a function of WIP and LMSF in no noise-condition and acoustically noisy condition. From the example, in the condition with the background babble noise, larger word insertion penalty (more negative) and language model scale factor help to suppress the insertion errors. The computation is carried out in the log domain to avoid underflow.

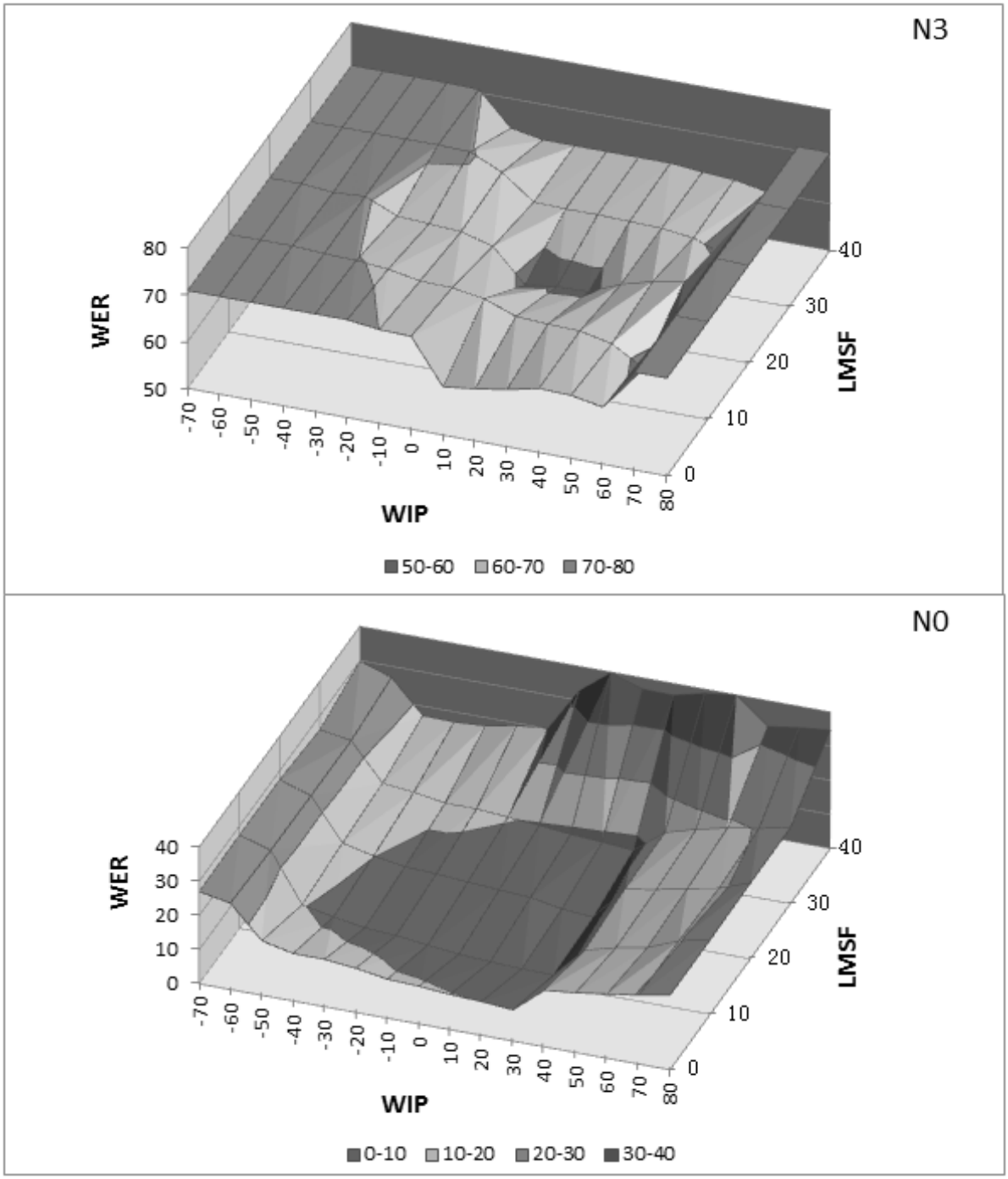


Figure 7.4: Surface plot showing WER as a function of word insertion penalty (WIP) and language model scale factor (LMSF) for the same speech recording with (N3, top) and without additive acoustic noise. Optimum values of WIP and LMSF are 30,10 for N3 and 10,20 for N0, showing that the larger penalty (more negative) and scale factor assist to counter balance the insertion errors brought by background babble noise.

7.3.5 LM performance measure

The performance of the language model is measured using perplexity (PER), which is calculated as the inverse of the geometric average probability assigned to a word sequence, $W = \{w_1, w_2, ..., w_N\}$, of N words:

$$PER = 2^{-\frac{1}{N}\log_2 P(W)} \tag{7.20}$$

A language model with a lower PER value is theoretically better at predicting the word probabilities in the test data. However, while the calculation of perplexity can be done without access to a speech recogniser, the PER value often does not correlate well with speech recognition WER [42]. Examples in the previous studies show that the language models with lower perplexity may yield little or no improvement in terms of WER in a real recogniser [191] [136].

Nevertheless, although perplexity does not necessarily represent the recognition performance, it is still an elegant and widely used measure for comparing language models with the same vocabulary [42]. Therefore, in this evaluation, the performances of the language models will be compared using the perplexity, and their recognition performances will be evaluated using a real ASR system.

7.3.6 ASR performance measure

A standard metric for ASR performance is word error rate (WER). WER is derived from the Levenshtein distance [177] considering the errors caused by substitution, deletion and insertion to have the equal weight. It can be computed as:

$$WER = \frac{N_s + N_d + N_i}{N} \tag{7.21}$$

where N is the total words number, N_s the number of substitutions, N_d the number of deletions, and N_i the number of insertions.

WER was used as a performance measure in the previous studies utilising gaze to improve speech recognition [257] [49]. In this evaluation, the WER performances of the adapted ASR will be compared.

7.3.7 N-best rescoring

An ASR generates the most likely word sequence using the Viterbi decoding. The WER is calculated by comparing this sequence with a reference sequence. With N-best rescoring, instead an ASR generates N hypothesised competing word sequences with a probability assigned to each of them. The so-called 'N-best list' is a list with these sequences listed according to their probabilities order. This list can be re-ordered combining the information from the adapted LM to account for the extra context awareness; in this case, the information from gaze. The updated probability $P^*(W)$ of word sequence W with L words is calculated as:

$$P^*(W) = P(W) \prod_{i=1}^{L} \frac{p_a(w_i|w_{i-1})}{p_b(w_i|w_{i-1})} \tag{7.22}$$

where $p_a(.)$ denotes the probabilities provided by the adapted LM $P_a^t(.)$ and $p_b(.)$ provided by the baseline LM $P_b(.)$ (see expression 7.12). As the WER for the N-best list is defined to be calculated using the entry with the highest probability score, the effect of the adapted LM can be assessed by comparing the WER of the rescored list.

It is a standard approach in ASR studies to use N-best rescoring to obviate the computational needs of Viterbi decoding. It is used as an approximation of Viterbi decoding with dynamic LM probabilities. In this work, to evaluate the LM adaptation in an ASR system, the LM probabilities for N-best rescoring are updated using the cache-based adapted LMs (see section 7.2.2). With a longer list (larger N), a better WER performance is expected because there will be more accurate entries present in the list. The choice of N is subject to the computational efficiency, accuracy and the size of vocabulary.

7.4 Tests Conducted

Four tests are conducted to evaluate the value of ANI and the selective use of gaze information with VAI to improve the ASR performance in noisy environments by cache-based class language model adaptation. The systems are validated with 7-fold cross validation with each fold contains data from a participant to ensure the test set are disjoint with the training sets.

7.4.1 Test 1: language model (LM) adaptation performance

The objective of the test is to evaluate the three LM adaptation approaches (see section 7.2.3) in terms of the perplexity improvement. It is hypothesised that the selective use of gaze based on the relevance function out-performs the non-selective use approach.

The baseline LM is constructed as discussed in Section 7.3.2. The perplexity of the baseline LM is 12.41.

The relevance score s calculated by VAI (see section 7.2.4) is used to adapt the LM. The varying interpolation weight λ and the performances of $\sigma(g, \omega)$ in the three LM adaptation approaches are compared. The results are reported in section 7.5.1 as perplexity.

7.4.2 Test 2: acoustic model (AM) adaptation performance

The objective of the test is to evaluate the optimum acoustic model adaptation technique and to demonstrate the value of the noise condition inference. The baseline acoustic model (referred to as BAM, see section 7.3.3) is adapted to no-noise (N0) speech (referred as N0AM) and noisy speech (N1, N2, and N3 as N1AM, N2AM, and N3AM). The N0AM model is compared with the N1AM, N2AM, and N3AM in terms of the recognition performance on the corresponding noisy speech. It is hypothesised that the performance of recognising noisy speech using the corresponding adapted acoustic model is better than using N0AM non-selectively (i.e., ASR performance can be improved by noise condition

inference) due to the reduced mismatch between the training and the recognition condition.

As discussed in section 7.1.6, the adaptation techniques MLLR, MAP, and the serialisation of them is used to adapt the BAM to the ES-N corpus data for each noise condition. The tests are performed following the procedure below, and the results are compared in terms of word error rate (WER).

- No adaptation - Baseline performance
- MLLR with a global mean and diagonal covariance transformation - Global MLLR
- MLLR with mean and diagonal covariance transformations using a regression tree - Regression MLLR
- MAP adaptation
- Regression MLLR followed by MAP

The adaptation data used here is the speech from all recordings, so the results are optimistic for the comparison between AM adaptation techniques. However, in the ASR evaluation, 7-fold cross validation are used with each fold contains data from a participant. The results are reported in section 7.5.2. The acoustic models adapted using the optimum technique (N0AM, N1AM, N2AM, and N3AM) are used in the next test.

7.4.3 Test 3: VAI-based LM adaptation performance in ASR

It is discussed in section 7.3.5 that the perplexity improvement does not correlate well with the speech recognition WER improvement. Thus, to evaluate the LM adaptation in terms of the WER improvement, the three LM adaptation approaches are tested in the ASR system. It is hypothesised that the ASR system benefits from the LM adaptation approach that uses gaze selectively (i.e., using VAI).

Evaluated on the speech data for each noise condition, the N-best rescoring technique (see section 7.3.7) is used to update the recognition results. The baseline ASR system

uses the optimum acoustic models in test 2 and the baseline LM in test 1. For the coupling function within VAI, the word sequence obtained by ASR and the recorded screen-display-change events are used. In section 7.5.3, WERs are measured to compare the performances of three LM adaptation approaches.

7.4.4 Test 4: selective gaze-Contingent ASR performance

The objective of the test is to evaluate the selective gaze-contingent ASR incorporating both VAI and ANI frameworks. It is hypothesised that the selective gaze-contingent ASR outperforms a baseline ASR system using a fixed acoustic model (N0AM) with the baseline LM.

The way to integrate ANI in the ASR system is that, for all noise conditions, the acoustic models were adapted beforehand and when a noise condition is inferred by the ANI, the corresponding model is selected for the recognition.

Evaluated on the aggregated/mixed speech data recorded in different noise conditions using the optimum approach in test 2 and 3, the ASR performance adapting VAI and VAI + ANI is reported in WER in section 7.5.4. The performance of ANI using the MI approach is more favourable than using other gaze and speech features (see section 5.6.5). However, it is away from the upper-bound (100%) performance. Because there are only 4 noise conditions and, ideally the adaptation is based on continuous acoustic noise inference, the ASR performance adapting VAI + upper-bound ANI is also reported.

7.5 ASR Results

7.5.1 LM adaptation perplexity performance (Test 1)

The cache length for gaze events is fixed at $l = 15$, which is determined empirically from values ranged from 1 to 30 to ensure the optimum average perplexity of the selective approach. For each task recording, adapted LMs with three implementations of $\sigma(g, \omega)$

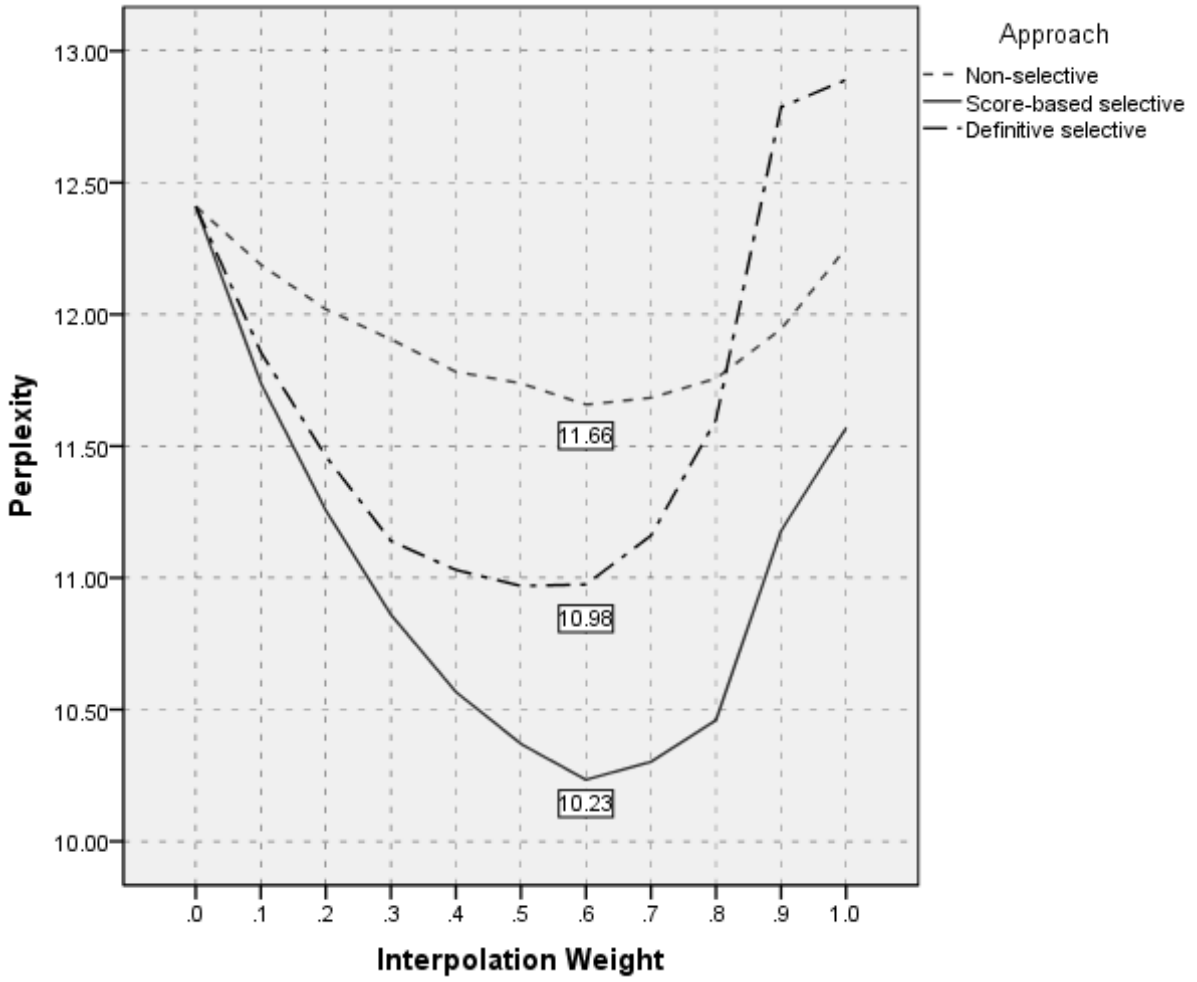


Figure 7.5: The perplexity values of the three approaches to relevance function for different LM interpolation weights λ. The score-based selective use of gaze outperforms definitive and non-selective use.

are compared with the baseline LM results. The interpolation weight λ ranges from 0 to 1 indicating the percentage contribution from the gaze modality (see expression 7.12), i.e., 0 means using 100% of the LM constructed from the speech information, and 1 means 100% of the LM derived from gaze information. The percentage improvements of the LM perplexity from the baseline LM are shown in Figure 7.5 and Table 7.1.

Using either class distribution alone ends up with a higher perplexity; lower perplexity is obtained when combining both models. This suggests that the language used in the task is better modelled when two LMs contribute approximately equally. When only using the baseline LM constructed from speech information, the statistics are drawn from a big set therefore lack more specific knowledge of a particular task recording. Conversely, using

	Perplexity					
	Non-S		Score-based S		Definitive S	
Weight	Mean	Std	Mean	Std	Mean	Std
0	12.41	1.69	12.41	1.69	12.41	1.69
0.1	12.19	1.68	11.74	1.60	11.86	1.62
0.2	12.02	1.68	11.26	1.55	11.46	1.58
0.3	11.91	1.69	10.86	1.52	11.14	1.56
0.4	11.78	1.71	10.57	1.51	11.03	1.56
0.5	11.74	1.73	10.37	1.53	10.97	1.57
0.6	11.66	1.77	10.23	1.57	10.98	1.60
0.7	11.68	1.82	10.30	1.66	11.16	1.66
0.8	11.76	1.89	10.46	1.82	11.60	1.78
0.9	11.94	1.99	11.18	2.21	12.79	2.04
1	12.25	2.15	11.57	2.51	12.89	2.10

Table 7.1: The perplexity of non-selective, score-based selective and definitive selective approaches. The score-based relevance function shows the lowest perplexity value, with best performance at $\lambda = 0.6$ (10.23).

only the information from gaze is less reliable and suffers from a sparseness problem as the VA-cached LM only contains words related to visual fields.

From Figure 7.5, it can be seen that when $\lambda = 0.6$, score-based approach (10.23, 17.57% improvement from the baseline performance) outperforms definitive approach (10.98, 11.52%) and non-selective approach (11.66, 6.04%) in terms of the perplexity improvement ($p < 0.01$, two-tailed t-test, same tests conducted after). When using the information from gaze only ($\lambda = 1$), the score-based selective approach also models the language better (11.57) than the other two approaches (12.25 and 12.89). This supports the expectation of considering the relevance prior to the use of gaze information is beneficial for the language model adaptation.

7.5.2 AM adaptation WER performance (Test 2)

The adaptation results are shown in Table 7.2 measured by WER. The WIP and LMSF are optimised for each setting. A regression tree with 32 terminal or leaf nodes was created using HTK for the transforms [325]. For all noise conditions, the ASR performs best with regression MLLR followed by MAP. Therefore MLLR followed by MAP is used

WER(%)	N0	N1	N2	N3
Baseline	88.79	91.83	92.02	93.35
G-MLLR	32.70	47.60	52.69	77.38
R-MLLR	23.33	34.55	51.57	70.61
MAP	65.98	87.50	87.88	93.87
MLLR+MAP	10.84	26.32	40.95	63.42
N0AM		80.05	94.48	98.15

Table 7.2: The adaptation results measured in WER. G-MLLR stands for the global MLLR and R-MLLR the regression MLLR.

as a benchmark performance in the next test with the adapted LMs. The corresponding performances of the ASR system are 10.84% in N0, 26.32% in N1, 40.95% in N2, and 63.42% in N3. The performance differences between the noise levels meet the purpose to distinguish the ASR to test the value of selective use of gaze respectively. This again justifies the choice of acoustic noise levels used in this study (section 4.5.1).

Acoustic model adaptation reduces the mismatch between the training and recognition conditions thus leading to a better ASR performance (see section 2.3.4). Using N0AM (see section 7.4.2) non-selectively on the noisy speech data has significantly worse performances compared to using the correspondingly adapted acoustic models due to the greater mismatch. The results demonstrate the potential benefit of noise condition inference.

The performances of MLLR and MAP adaptations can vary depending on the size of the training data size. For the MAP to perform better, normally a large amount of adaptation data is required. As the Regression MLLR adaptation (65.46% improvement from baseline in N0, 57.28% in N1, 40.45% in N2, and 22.74% in N3) performs better than the global MLLR (56.09% improvement from baseline in N0, 44.23% in N1, 39.33% in N2, and 15.97% in N3). It indicates that there are enough data for forming a regression tree.

In regard to MAP, although it shows considerable improvement after the adaptation (22.81%) in N0, the lower performance comparing to the MLLR indicates that the amount of adaptation data is not enough for MAP to beat MLLR. In N3, a plausible reason for the less favourable performance of MAP can be that the babble noise sounds as many

people talking. Thus, the adaptation result is greatly compromised as MAP primarily accounts for speaker variation.

Compared to the 20.9% benchmark performance against the WSJCAM0 test data (section 7.3.3), the good recognition performance in no noise condition can be related to the fact that users in the 'put-that-there' task with gaze and speech as inputs tend to use short commands [75]. Note that the LM used in the ASR for the acoustic adaptation tests is the baseline LM described in section 7.3.2.

7.5.3 ASR WER performance for evaluating the LM adaptation (Test 3)

Figure 7.6 and 7.7 illustrate the ASR performances measured by the N-best rescoring using the adapted LMs in no noise (N0) and most noisy condition (N3) respectively. It can be noted that the decrease of WER slopes gently for $N > 100$ in both conditions. Consequently, the 100-best list results listed in Table 7.3 are reported as the ASR performance measurement.

In no-noise condition, the non-selective $\sigma(g, \omega)$ LM adaptation has a 3.09% ($p < 0.01$, two tailed t-test, same tests conducted after) absolute improvement in WER from no LM adaptation. This is comparable with the previous studies [49] [257]. In acoustically noisy conditions, this improvement rises to 8.46% in N1, 10.2% in N2 and 12.29% in N3, rendering the increasing value of using gaze information to assist speech recognition in acoustic noise. The selective use of gaze shows a further improvement of 3.14% ($p < 0.001$) in N2, and 5.69% ($p < 0.001$) in N3, while no significant difference is reported between no-noise (N0) and lower noisy conditions (N1, $p > 0.05$). The results demonstrated that the ASR benefits more from using gaze selectively in louder noise. Meanwhile, no difference ($p > 0.1$) is found between the two selective approaches.

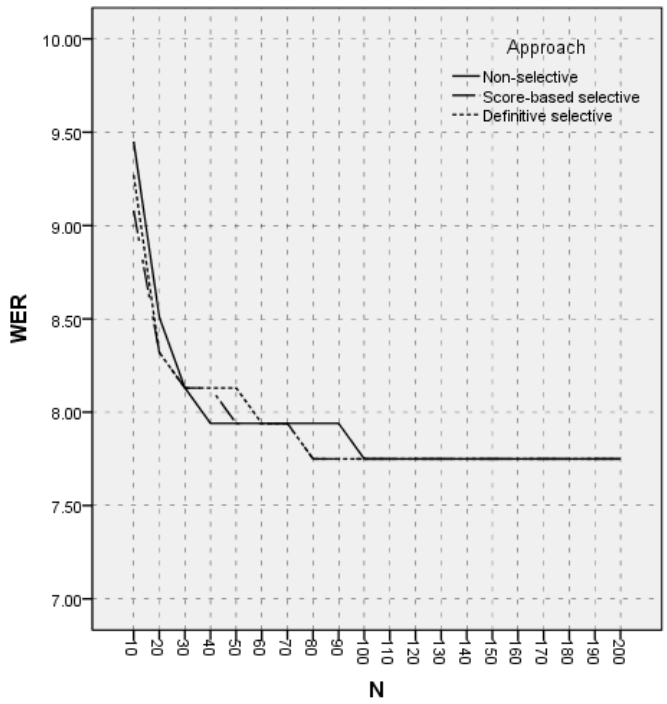


Figure 7.6: The figure shows the effect of LM adaptation in no-noise condition (N0). WER differs as the length N of N-best list changes.

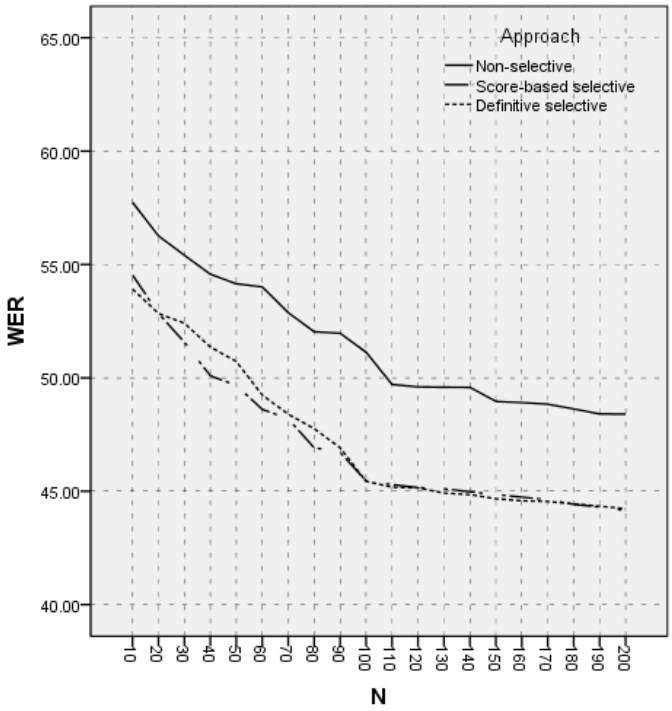


Figure 7.7: The figure shows the effect of LM adaptation in the noisiest condition (N3). WER differs as the length N of N-best list changes.

WER(%)	N0	N1	N2	N3
Before LM adaptation	10.84	26.32	40.95	63.42
Non-selective	7.75	17.86	30.75	51.13
Definitive selective	7.75	16.98	27.61	45.44
Score-based selective	7.75	16.96	27.40	45.45

Table 7.3: The WER performances before and after the LM adaptations. $N = 100$ is used in the N-best list. The selective use of gaze is more valuable in the noisier condition.

7.5.4 Selective gaze-contingent ASR WER performance (Test 4)

Based on the test results above, the selective gaze-contingent ASR employs ANI with the MLLR+MAP adaptation technique and VAI with the score-based LM adaptation approach. The complete system is evaluated using the mixed speech data recorded in different noise conditions (N0, N1, N2, and N3). While in Test 2 and Test 3, all data are used to achieve the optimistic results for comparison, in this test the 7-fold cross validation is employed with each fold contains data from one participant. To demonstrate the noise-robustness of the selective gaze-contingent ASR, the baseline ASR adapted to the no-noise ES-N speech (N0AM) is used for the benchmark performance. Table 7.4 shows the breakdown of the baseline performance in terms of noise conditions and participants.

Baseline (N0AM)		N0	N1	N2	N3
Participant	1	16.3	77.11	91.03	98.91
	2	8.7	78.85	85.37	95.45
	3	2.44	92.5	97.83	95.27
	4	11.29	95.89	90.8	95.55
	5	4.11	75	89.23	97.1
	6	16.05	88.89	93.33	94.5
	7	12.04	85.71	94.05	97.7
	Overall	10.13	84.85	91.66	96.34

Table 7.4: The breakdown of the baseline performance in terms of noise conditions and participants.

Table 7.5 reveals that the benchmark WER of 73.01% is improved by 2.68% (absolute improvement, same for the following improvements) using information from gaze with ANI. Compared with using ANI alone, using VAI yields more improvement (14.73%) due to its better overall performance than the former. Incorporating coupling function boosts

VAI's performance by 2.08% in the ASR system. For demonstration, when oracle speech instead of the output from the baseline ASR is used in the coupling function, the performance is further improved by 6.23%. The actual system (ANI+VAI) improves the ASR performance by 19.02% and 6.97% further than using ANI and VAI alone respectively. For completeness, the hypothetical upper-bound ANI assuming 100% ANI accuracy is also included in the table. This result stands for the ideal situation where the noise condition can always be acknowledged so the corresponding acoustic model can always be applied. Even when the upper-bound ANI is employed, incorporating VAI would further yield an improvement by 6.4%, indicating the benefit of the selective use of gaze information in the ASR system. It can also be noted that using gaze information via ANI together with VAI benefits the speech recognition for all participants.

	Participant							
	Overall	1	2	3	4	5	6	7
Baseline	73.01	73.11	69.39	74.39	75.42	68.92	75.2	74.63
ANI	70.33	68.14	66.63	70.15	73.09	68.63	75.13	70.53
VAI (without coupling funtion)	60.36	61.76	59.47	56.13	63.39	55.8	62.11	63.73
VAI	58.28	56.48	52.45	62.38	60.88	53.8	59.18	62.78
VAI(oracle speech coupling)	52.05	48.57	51.45	55.48	51.4	49.1	56.02	52.35
VAI+ANI	51.31	45.9	50.16	53.67	53.86	50.13	54.04	51.41
Upper-bound(100%) ANI	37.18	37.86	34.79	40.29	36.94	37.09	38.5	34.82
VAI + upper-bound ANI	30.78	30.61	31.37	26.07	30.68	29.42	32.29	35.05

Table 7.5: ASR system performances in terms of WER on gaze and speech data recorded in various acoustic noise.

7.6 Summary

In this penultimate chapter, gaze is successfully used selectively to improve ASR performance. The selection are based on VAI and ANI.

In Chapter 5, the hypothesised cognition roles for gaze are described and its relationships with speech are quantified for the inference of acoustic noise. The inference results are used for the ASR to employ corresponding counter-noise strategies. In this work, MAP and MLLR adaptation techniques are used for the noise condition demonstration.

In Chapter 6, the interaction and environmental reaction roles for gaze are inferred from classifiers trained using supervised learning. VAI has been demonstrated to be capable of distinguishing the speech-relevant gaze events by investigating the gaze characteristics and the multimodal relationship with other modalities. It is used in the LM adaptation. A cache-based class language model adaptation framework is represented using the VAI results to improve ASR performances. Instead of the commonly used words history, the adaptation incorporates a cache composed by gaze events. The information from gaze events is used selectively via relevance functions based on the fact that not all gaze events are contributing towards the system function (e.g., to assist speech). Class-based language model is used with the prior knowledge of task to reflect the speech-specific relation between words and visual foci.

The LM adaptation results show the improvements over the baseline LM by adapting the gaze information measured by perplexity. The optimum interpolation weight λ is investigated, and for all λ values, the score-based selective approach has the most notable improvements. The perplexity serves as a measurement to compare LM performance. The recognition performances are evaluated using an ASR.

An ASR trained with the WSJCAM0 corpus for a speech-gaze map task is used as the baseline for the evaluation of the framework. The regression MLLR followed by MAP gives the lowest WER and is therefore used as the benchmark for evaluating the recognition performances by adapting LM.

N-best list rescoring of ASR output is used to measure the recognition performance of adapted LMs. The list length of 100 is chosen empirically considering the computational efficiency, accuracy, and vocabulary size. Statistically significant improvements in WER are demonstrated with the greatest improvement in the acoustically noisiest condition - the scenario where using information from gaze is expected to be more beneficial. The selective use of gaze illustrates more favourable results to the non-selective approach in acoustically noisy conditions, validating the value of VAI. Although the score-based LM is illustrated to be more desirable than the definitive LM, no significant difference has

been found in terms of their WER performance.

With the optimum acoustic model adaptation technique (MLLR+MAP) and language model adaptation approach (score-based) tested, the selective contingent-gaze ASR is evaluated on the mixed data recorded in different noise conditions. The results reveal the promising value of using gaze selectively considering its role and relationship with speech via VAI and ANI framework in the ASR systems.

In the final chapter, the contributions of the thesis are highlighted, and the recommendations for the further work are proposed.

CHAPTER 8

CONCLUSION

In Chapter 1, the following questions are posted related to the use of gaze selectively for the integration with speech to improve ASR noise-robustness:

- How to integrate the information events in gaze and speech considering their relationship (temporal and semantic)?
- Is gaze's behaviour and relationship with speech dependent upon acoustic noise? Can this dependency be exploited for ASR?
- How to use gaze selectively to integrate with speech by considering its relevance?

A formal framework for multimodal coupling is proposed for the integration with speech using gaze selectively. To implement and evaluate the framework, an eye/speech corpus is collected in different noise conditions, and a research-level ASR with a task-specific language model is built and tested on the data collected. Information from gaze and speech is sensed and characterised with their relationship dependency upon acoustic noise investigated. A taxonomy of gaze roles is proposed considering the underlying cognitive, interactional, and environmental aspects of context awareness. Different gaze events are used selectively based on their hypothesised roles and measurability in HCI systems for the ASR performance improvement. The system is evaluated on the data recorded in acoustically noisy environments, and recognition performances are compared.

From the aspect of multimodal HCI system engineering and ASR research, this thesis discusses the motivations and the related previous studies, describes the frameworks and the implementations, comments on the results and suggests enhancement. In this chapter, the major contributions are highlighted and the suggested directions for future researches are discussed.

8.1 Contributions

The thesis addresses the research questions with a novel application of integrating gaze selectively into an ASR system and the evaluation on data recorded in acoustically noisy environments. In this section, the major contributions are summarised.

8.1.1 A formalised framework for measuring the coupling between modalities

To address the coupling between information events in loosely coupled (i.e. semantically asymmetrical and temporally asynchronous) modalities such as gaze and speech, a coupling framework is proposed. Instead of using the traditional model which only couples a speech event with either the preceding or the co-occurring gaze event, the speech event can be coupled with any gaze event by a coupling strength function composed of a semantic and a temporal component. In doing so, the framework can better accommodate human's natural gaze-speech behaviour patterns (i.e. does not force the user to use gaze deliberately). The implementations of the general framework are proposed based on the hypothesised roles of gaze. This framework can be applied to model the coupling between speech and other loosely coupled modalities, such as body gestures.

8.1.2 A working taxonomy of gaze roles

The concept of gaze role measurability is outlined considering the cognitive and interactional meanings of gaze. A gaze role taxonomy is described to distinguish visual attention types with the underlying aspects of context awareness. The gaze roles are distinguished by whether they can be validated (i.e., directly measured), and two corresponding frameworks are developed for the selective use of gaze. The use of these two frameworks is highlighted for the improvement of ASR noise-robustness. Their implementations are presented and evaluated on the collected corpus data and applied to a real ASR system. The taxonomy is useful for gaze researches to explicitly account for roles rather than to assume unnatural constraints on user behaviour or to model variation in gaze behaviour with random variables.

8.1.3 The ES-N corpus recorded in different acoustic noise

A corpus of matched gaze and speech data is collected within a task inspired by 'put-that-there'. To account for the user's behavioural changes, the corpus is recorded in different acoustically noisy environments compared to most other studies where the noise is used to contaminate the data recorded in a clean environment. The task is designed based on the 'Wizard of Oz' simulation paradigm to avoid the expensive cost and the technical shortfall in building a real intelligent system that can robustly perceive and understand the natural gaze and speech behaviours of humans. The paradigm also allows quick set-up of a pilot study for the designer to investigate the experiment apparatus, types of data to collect, and the acoustic noise type to use. The hardware and software platform allows the data to be recorded synchronously. Speech recorded using desk-mounted microphones is transcribed and time-aligned. Eye gaze is recorded using a head-mounted eye-tracker, and the quality is assessed both subjectively and objectively. The gaze and speech data is annotated for further analysis. The system responses are also recorded for the visual attention inference. It highlights the need for multimodal evaluations to consider acoustic

noise as a variable.

8.1.4 The 'gaze Lombard effect' and the dependency of the gaze-speech relationship on acoustic noise

Systemic statistical analyses are performed on the speech and gaze data for their change in acoustic noise and variability between people. The speech analysis supports the previous findings of acoustic Lombard effect. The analyses in 'gaze Lombard effect' reveal the change of fixation durations and saccade lengths in acoustic noise and the variability between people. The dependency upon acoustic noise for fixations 'during speech' and 'during silence' is investigated in addition to their relative changes. The semantic and temporal relationship between gaze and speech is quantified based on the coupling framework described using information theoretic measures based on mutual information. The dependency of the relationship upon noise is explored, and the results are utilised for the inference of noise condition in the ASR. A SVM classifier is built to compare the MI measure to speech and gaze characteristic features in terms of the discriminability of noise condition and variability between persons. Based on the finding, it is reasonable to anticipate that the Lombard Effect can exist in other non-verbal modalities, such as gesture.

8.1.5 A cache-based LM adaptation approach using class-based model

For the on-line integration of gaze events into speech recognition, a cache-based language model adaptation framework is proposed with the cache composed by gaze events. A class-based bi-gram model is constructed with the classes representing different groups of speech-related visual objects. Relevance functions are formalised for the information from these gaze events to be used selectively based on the fact that not all gaze events are contributing towards the assistance of speech comprehension. A visual attention inference framework is proposed based on the gaze roles associated with the interaction

and the reaction to the environmental changes. The framework that involves using a naive Bayesian classifier is applied specifically to implement the relevance functions using three different approaches. The approach highlights that the use of non-verbal modalities in LM adaptation be selective based on the relevance to speech.

8.1.6 A noise-robust, selective gaze-contingent ASR

With the noise inference and visual attention inference frameworks evaluated respectively, a novel gaze-contingent ASR is constructed that incorporates gaze selectively based on these two frameworks. A task-specific class-based baseline language model is constructed to interpolate with the cache-adapted language models. An ASR trained with the WSJ-CAM0 corpus for a speech-gaze task is used as the baseline for the evaluation of the framework.

MLLR and MAP acoustic model adaptation techniques are used as counter-noise strategies in this study for the demonstration of incorporating noise inference framework into the ASR system. The performances of the adapted language models utilising the visual attention inference framework are measured using perplexity for the comparison of three adaptation approaches. The adapted language models are then tested in the ASR system with the performances measured in word error rate. With the optimum acoustic model and language model adaptation approach tested, the gaze-contingent ASR that realises the selective use of gaze by incorporating noise inference and visual attention inference frameworks is evaluated on the mixed speech data recorded in various noise conditions and the noise-robust performances are reported. The work demonstrates the value for a methodical use of non-verbal modalities that is loosely coupled with speech, such as gaze, in both acoustic and language model adaptations to improve ASR noise-robustness.

8.2 Recommendations for Future Research

The theoretical and technical contributions towards the selective use of gaze for the integration with speech provide the following insights into the future researches:

8.2.1 Corpus and General Framework

The ES-N task used in the study is designed specifically to address the cooperative use and comprehension of gaze and speech for object manifesting in multimodal HCI systems. The general coupling framework for loosely coupled modalities is evaluated in such context. However, whether the framework can be applied on other modalities such as gestures and body movements needs the corresponding tasks to be set-up and data to be recorded. The premise that inferring gaze roles related to cognition is not possible may be challenged by advances in brain scanning. However, inferring cognition from the connectionist structure of the brain in real time and non-intrusively is not feasible with current sensing technologies and require future developments in the related hardware and software. The approach outlined in this thesis can be achieved with current technologies and has the potential to work in unison with evolving brain sensing technologies.

Resources available for this study enabled the ES-N corpus to be recorded with the relative small number of participants. However, any multimodal study would benefit from capturing larger data sets from more participants (e.g. > 30). If consider the possible personal dependency and variations in speech and gaze sensing such as gender, region, accent, lenses wearing and prescriptions etc., a much larger participant group would improve the system reliability and user-independency. However, the resource and time need for such a large corpus is beyond the scope of current study. A larger corpus would be useful in supporting the findings in this study or incorporating extra modalities. The head-mounted eye-tracker used in this study can be intrusive and restrict the users' natural head movement. It may also cause calibration errors and result the recording session to be discarded. This can be solved by the recent advances in miniaturising eye-

tracking technologies.

Although the VAI evaluation assumes implementations for the main classes (TOVA, TIVA and RVA), the common approach highlights the framework's potential generalisability to different system task assumptions. Future HCI systems could have a gaze role inference function utilising Context Awareness from a cognitive, environment and interactional perspective. A design that builds upon but ultimately supplants natural human-to-human interaction could infer a person's thoughts and experience. The roles described in this work are not gaze-specific but could be applied to other non-verbal modalities.

8.2.2 Noise Condition and Affective State

To account for the effect of acoustic noise, the data in this study is recorded in four different noise conditions. The noise type used is the non-stationary multi-speaker babble noise from NoiseX-92 corpus which amplified to different noise levels. The effect of the noise is revealed to be dependent on the noise type and loudness level. Thus it is expected the use of more noise condition is useful in supporting the findings in this study or developing noise-dependent strategies for the further improvement. For example, although results for four noise conditions are reported in this study and a positive correlation has been found between the noise level and the value of using gaze selectively, the exploration of whether such correlation only exists below certain noise level remains for the future work. Also, the different types of noise may have other impacts on the users' behaviours and recognition performance. However, considering the noise types and levels presented in the real-world environments, it is very unlikely that exhaustive acoustic noise conditions can represent all different scenarios. Nevertheless, it is still likely that with more noise conditions explored in the future researches a relatively reliable correlation can be formalised between the acoustic noise and the human behaviour changes.

In the recording, noise levels were changed in a fixed order; N0, N1, N2, and N3. There is a possibility that different orders of noise levels may elicit different behaviours. In this

study, there was a break between noise levels. Task performance which can be affected most is the trial length and it is not used in building the system. In future study, the effect of the noise order can potentially become a research topic.

Besides the acoustic noise, other forms of variations such as lighting conditions can also change people's interactional behaviours. In addition to the external noises, internal affective state has also been demonstrated to affect the interaction [326]. These aspects are not involved in this thesis but for a machine to be fully capable of understanding a human's natural interactional behaviours, they need to be addressed in the future researches.

8.2.3 Acoustic Noise Inference Using Statistical Gaze Information

The paper aims to exploit the selective use of gaze behaviours in speech recognition. It demonstrates the possibility to use gaze in summary-based approach (ANI) and also event-based approach (VAI). ANI, as a component of the selective gaze-contingent ASR, exploits the 'gaze Lombard effect', and provides a possibility to use summary-based gaze metrics and its relationship with speech (MI) in noise classification. Function-wise, it can be replaced by a more robust speech-only approach but that would not be the focus of this explorative work.

Lombard effect was normally reported as summary statistics. In order to compare other feature sets with MI measure, which is a summary-based metric, SVM was used to process the mean values.

As an innovative research which tries to exploiting different gaze roles during the interaction, the approach aims to shed some light for the researches in the area. Together with the event-based approach in VAI, this summary-based approach in ANI is discussed for completeness and providing a possible means of using gaze related to cognition inference. The future advance in cognition-recognition technology (e.g., brain scanning) may benefit this process. Although far from robust yet, the work demonstrate a possible way of using gaze in ASR systems which can be potentially improved, or be generalised to

other loosely coupled modalities in the future researches.

8.2.4 Lombard Effect in Different Modalities and Variability between People

The 'gaze Lombard Effect' and the acoustic Lombard Effect are investigated in this work, with the latter supporting the previous findings. Variability is reported to exist between people. A lot of previous studies have reported the between-people variations in the acoustic Lombard Effect with the explanation given as the difference in personal condition or habit. An interesting observation is reported in section 5.5 which shows a correlation between fixation duration and speech power changes. Thus, a likely interpretation of the variability is given as the person's freedom of enhancing intensity in different modalities to increase the communication intelligibility in acoustic noise. However, confident conclusions need to be made with more data and more specific experiments that possibly monitor more modalities.

That the noise is played to the wizard is because the purpose of the WoZ paradigm is to simulate the real system where noise is involved so the users will behave correspondingly. It is possible that the misrecognition of wizard can impose certain behaviour change of the user. However, the misrecognition rate is very low (2.9%) in this case which is very unlikely to be the cause of the overall 'gaze Lombard effect' across all sessions.

8.2.5 ASR Vocabulary, Segmentation and Language Model

Because the speech data is collected in a task-specific HCI experiment, the vocabulary is limited compared to the continuous speech corpora designed for more general use. With a larger and more general gaze speech corpus recorded in noise environments, the selective gaze-contingent ASR can be validated with a less specific vocabulary. The word segmentation model used in this ASR system is a silence model. Replacing it with a more noise-robust segmentation approach may result a better performance. In this work,

bigram language model is used for the framework validation due to its ease of use and vocabulary size. Other models like trigram can be used in the future for the potential further performance improvement.

The participants in the experiment are non-local speakers. Generally speaking, accent can degrade the ASR performance. In this study, the results are considered not compromised due to the following facts. First, the vocabulary of the task is relatively small. The instruction commands are very short and the grammar is easy and straight-forward (e.g., Blue circle on top). Then, all participants are from the same country and they are all highly-educated PhD student in UK. And acoustic model adaptation is performed in all cases.

8.3 Summary

Compared to the number of the studies in multimodal systems using gaze and speech inputs, the volume of the studies that use gaze specifically to improve ASR performance is very limited. Although some of them have shown improvement in ASR performance, none of these reported studies were evaluated on data recorded in acoustically noisy environments, or considered the selective use of gaze based on different roles which are distinguishing between those which are measurable (interaction) and not (cognition).

Noise-robust ASR requires a variety of strategies. A methodical use of information from non-verbal modalities, such as gaze, into both acoustic and language models shows promise.

APPENDIX A

PUBLICATION

Two published conference paper are listed.

- Ao Shen/ Neil Cooke/ Martin Russell (2013): 'Selective use of gaze information to improve ASR performance in noisy environments by cache-based class language model adaptation.', in INTERSPEECH-2013.
- Ao Shen/ Neil Cooke/ Martin Russell (2014): 'Exploiting a 'gaze-Lombard effect' to improve ASR performance in acoustically noisy settings.', in ICASSP 2014

LIST OF REFERENCES

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *Handheld and ubiquitous computing*, pages 304–307. Springer, 1999.

[2] Antti Ajanki, David R Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. Can eyes reveal interest? implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*, 19(4):307–339, 2009.

[3] Patrice Alexandre and Philip Lockwood. Root cepstral analysis: A unified view. application to speech processing in car noise environments. *Speech Communication*, 12(3):277–288, 1993.

[4] David B. Allison, Bernard S. Gorman, and Elizabeth M. Kucera. Unicorn: A program for transforming data to approximate normality. *Educational and Psychological Measurement*, 55(4):625–629, 1995.

[5] Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

[6] Giuliano Antoniol, Roldano Cattoni, Mauro Cettolo, and Marcello Federico. *Robust speech understanding for robot telecontrol*. Istituto per la Ricerca Scientifica e Tecnologica, 1993.

[7] MA Anusuya and Shriniwas K Katti. Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*, 2010.

[8] Michael Ashmore, Andrew T Duchowski, and Garth Shoemaker. Efficient eye pointing with a fisheye lens. In *Proceedings of Graphics interface 2005*, pages 203–210. Canadian Human-Computer Communications Society, 2005.

[9] Bishnu S Atal and Suzanne L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50:637, 1971.

[10] BS Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55:1304, 1974.

[11] Richard W Backs and Larry C Walrath. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied ergonomics*, 23(4):243–254, 1992.

[12] Brian P Bailey and Shamsi T Iqbal. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(4):21, 2008.

[13] Gérard Bailly, Stephan Raidt, and Frédéric Elisei. Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6):598–612, 2010.

[14] Jon Barker and Martin Cooke. Modelling the recognition of spectrally reduced speech. *cognition*, 12(9):U1, 1997.

[15] Jon P Barker and François Berthommier. Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*, 1999.

[16] Leonard E Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[17] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. In *Psychological Bulletin*, pages 276–292, 1982.

[18] Jackson Beatty and Brennis Lucero-Wagoner. The pupillary system. 2000.

[19] Gary G. (Ed) Beatty Jackson; Lucero-Wagoner. (Ed); Tassinary, Louis G. (Ed); Berntson. The pupillary system. In *Handbook of psychophysiology (2nd ed.)*, pages 142–162, 2000.

[20] Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. Eye-movements as a biometric. In *Image analysis*, pages 780–789. Springer, 2005.

[21] Roman Bednarik, Hana Vrzakova, and Michal Hradis. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 83–90, New York, NY, USA, 2012. ACM.

[22] Keni Bernardin and Rainer Stiefelhagen. Audio-visual multi-person tracking and identification for smart environments. In *Proceedings of the 15th international conference on Multimedia*, pages 661–670. ACM, 2007.

[23] Lynne E Bernstein, Marilyn E Demorest, and Paula E Tucker. *What makes a good speechreader? First you have to find one*. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998.

[24] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113–120, 1979.

[25] Hynek Boril and John HL Hansen. Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1379–1393, 2010.

[26] Hynek Boril and John HL Hansen. Ut-scope: Towards lvcsr under lombard effect induced by varying types and levels of noisy background. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4472–4475. IEEE, 2011.

[27] Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Lawrence D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, volume 2, pages 77–82. IEEE, 1994.

[28] Sahar E Bou-Ghazale and John HL Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 8(4):429–442, 2000.

[29] Norman Breslow. A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57(3):579–594, 1970.

[30] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377, 1996.

[31] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[32] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December 1992.

[33] F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–370, 1993.

[34] Andreas Bulling, Daniel Roggen, and Gerhard Troster. What's in the eyes for context-awareness? *Pervasive Computing, IEEE*, 10(2):48–57, 2011.

[35] Robert Burkard. Sound pressure level measurement and spectral analysis of brief acoustic transients. *Electroencephalography and Clinical Neurophysiology*, 57(1):83 – 91, 1984.

[36] Ellen Campana, Jason Baldridge, John Dowding, Beth Ann Hockey, Roger W Remington, and Leland S Stone. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM, 2001.

[37] Christopher S Campbell and Paul P Maglio. A robust algorithm for reading detection. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–7. ACM, 2001.

[38] Ruth Campbell, Barbara J Dodd, and Denis K Burnham. *Hearing Eye II: The Psychology of Speechreading and Auditory-Visual Speech*, volume 2. Psychology Pr, 1998.

[39] J. Carletta, R.L. Hill, C. Nicol, T. Taylor, J.P. de Ruiter, and E.G. Bard. Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior research methods*, 42(1):254–265, 2010.

[40] HE Çetingül, Engin Erzin, Yucel Yemez, and A Murat Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal processing*, 86(12):3549–3558, 2006.

[41] Lei Chen, R Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, et al. Vace multimodal meeting corpus. In *Machine Learning for Multimodal Interaction*, pages 40–51. Springer, 2006.

[42] Stanley F Chen, Douglas Beeferman, and Roni Rosenfield. Evaluation metrics for language models. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[43] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

[44] Y Chow, M Dunham, O Kimball, M Krasner, G Kubala, J Makhoul, P Price, S Roucos, and R Schwartz. Byblos: The bbn continuous speech recognition system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, volume 12, pages 89–92. IEEE, 1987.

[45] Chuang-Hua Chueh and Jen-Tzung Chien. Topic cache language model for speech recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5194–5197. IEEE, 2010.

[46] Phil Cohen, Colin Swindells, Sharon Oviatt, and Alex Arthur. A high-performance dual-wizard infrastructure for designing speech, pen, and multimodal interfaces. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 137–140. ACM, 2008.

[47] Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the fifth ACM international conference on Multimedia*, pages 31–40. ACM, 1997.

[48] Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 34(3):267–285, 2001.

[49] N. Cooke and M. Russell. Gaze-contingent automatic speech recognition. *IET signal processing*, 2(4):369–380, 2008.

[50] Neil J Cooke and Martin J Russell. Cache-based language model adaptation using visual attention for asr in meeting scenarios. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 87–90. ACM, 2009.

[51] Neil James Cooke. Gaze-contingent automatic speech recognition, 2006.

[52] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[53] Laura Cowen, Linden Js Ball, and Judy Delin. An eye movement analysis of web page usability. In *People and Computers XVI-Memorable Yet Invisible*, pages 317–335. Springer, 2002.

[54] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):30–42, 2012.

[55] N. Dahlback, A. Jonsson, and L. Ahrenberg. Wizard of oz studies–why and how. *Knowledge-based systems*, 6(4):258–266, 1993.

[56] Subrata Das, Raimo Bakis, Arthur Nádas, David Nahamoo, and Michael Picheny. Influence of background noise and microphone on the performance of the ibm tangora speech recognition system. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 71–74. IEEE, 1993.

[57] McNeill David. Hand and mind: What gestures reveal about thought, 1992.

[58] Chris Davis, Jeesun Kim, Katja Grauwinkel, and Hansjörg Mixdorff. Lombard speech: Auditory (a), visual (v) and av effects. In *Proceedings of the Third International Conference on Speech Prosody*, pages 248–252, 2006.

[59] Gillian M Davis. *Noise reduction in speech applications*, volume 7. CRC Press, 2002.

[60] KH Davis, R Biddulph, and S Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24:637, 1952.

[61] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[62] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[63] Heiner Deubel and Werner X Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.

[64] Gearge R Doddington and Thomas B Schalk. Computers: Speech recognition: Turning theory to practice: New ics have brought the requisite computer power to speech technology; an evaluation of equipment shows where it stands today. *Spectrum, IEEE*, 18(9):26–32, 1981.

[65] Martin J Doherty and James R Anderson. A new look at gaze: Preschool children's understanding of eye-direction. *Cognitive Development*, 14(4):549–571, 1999.

[66] John J Dreher and John O'Neill. Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America*, 29:1320, 1957.

[67] A.T. Duchowski. *Eye tracking methodology: Theory and practice*. Springer-Verlag New York Inc, 2007.

[68] Geoffrey B Duggan and Stephen J Payne. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1141–1150. ACM, 2011.

[69] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction*, pages 3–26. Springer, 2009.

[70] Kathleen M Eberhard, Michael J Spivey-Knowlton, Julie C Sedivy, and Michael K Tanenhaus. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6):409–436, 1995.

[71] James J Egan. *Psychoacoustics of the Lombard voice reflex*. PhD thesis, Western Reserve University, 1967.

[72] Shahram Eivazi and Roman Bednarik. Predicting problem-solving behavior and performance levels from visual attention data. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM IUI, 2011.

[73] Shahram Eivazi and Roman Bednarik. Predicting problem-solving behavior and performance levels from visual attention data. In *the proceedings of 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI*, volume 2011, pages 9–16, 2011.

[74] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, pages 57–106, 1981.

[75] Monika Elepfandt. Pointing and speech: comparison of various voice commands. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 807–808. ACM, 2012.

[76] Steve Ellis, Ron Candrea, Jason Misner, Christopher Sean Craig, Christopher P Lankford, and Thomas E Hutchinson. Windows to the soul? what eye movements tell us about software usability. In *Proc. 7th Annual Conf. Usability Professionals Association, Washington DC*, 1998.

[77] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109–1121, 1984.

[78] Yariv Ephraim and Harry L Van Trees. A signal subspace approach for speech enhancement. *Speech and Audio Processing, IEEE Transactions on*, 3(4):251–266, 1995.

[79] Lee D Erman, Frederick Hayes-Roth, Victor R Lesser, and D Raj Reddy. The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys (CSUR)*, 12(2):213–253, 1980.

[80] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *Neural Networks, IEEE Transactions on*, 20(2):189–201, 2009.

[81] Friedrich Faubel, Munir Georges, Kenichi Kumatani, Andrés Bruhn, and Dietrich Klakow. Improving hands-free speech recognition in a car through audio-visual voice activity detection. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 70–75. IEEE, 2011.

[82] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[83] JD Ferguson. Hidden markov analysis: an introduction. *Hidden Markov Models for Speech*, pages 8–15, 1980.

[84] John M Findlay and Iain D Gilchrist. Visual attention: the active vision perspective. In *Vision and attention*, pages 83–103. Springer, 2001.

[85] Dennis F Fisher, Richard A Monty, and John W Senders. *Eye movements: cognition and visual perception*. L. Erlbaum Associates, 1981.

[86] Paul M Fitts. Human engineering for an effective air-navigation and traffic-control system. 1951.

[87] Paul M Fitts, Richard E Jones, and John L Milton. Eye movements of aircraft pilots during instrument-landing approaches. *Ergonomics. 3. Psychological mechanisms and models in ergonomics*, 3:56, 2005.

[88] James W Forgie and Carma D Forgie. Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, 31:1480, 1959.

[89] Jesse Fox and Jeremy N Bailenson. Virtual virgins and vamps: The effects of exposure to female characters sexualized appearance and gaze in an immersive virtual environment. *Sex roles*, 61(3-4):147–157, 2009.

[90] J Friedman. Another approach to polychotomous classifcation. Technical report, Technical report, Stanford University, Department of Statistics, 1996.

[91] Georgios Galatas, Gerasimos Potamianos, and Fillia Makedon. Audio-visual speech recognition incorporating facial depth information captured by the kinect. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2714–2717. IEEE, 2012.

[92] Georgios Galatas, Gerasimos Potamianos, Alexandros Papangelis, and Fillia Makedon. Audio visual speech recognition in noisy visual environments. In *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*, page 19. ACM, 2011.

[93] Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[94] Mark John Francis Gales. Model-based techniques for noise robust speech recognition. 1995.

[95] Tanu Gandhi, Videep Kumar Antiwal, and Anuj Kumar Jain. Evaluation of smoothed language models. *International Journal of Research and Reviews in Computer Science (IJRRCS)*, 2(2), 2011.

[96] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.

[97] J-L Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing, IEEE Transactions on*, 2(2):291–298, 1994.

[98] Arne John Glenstrup and Theo Engell-Nielsen. Eye controlled media: Present and future state. *University of Copenhagen, DK-2100*, 1995.

[99] JH Goldberg and XP Kotval. Eye movement-based evaluation of the computer interface. *Advances in occupational ergonomics and safety*, pages 529–532, 1998.

[100] Joseph H Goldberg, Mark J Stimson, Marion Lewenstein, Neil Scott, and Anna M Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 51–58. ACM, 2002.

[101] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech communication*, 16(3):261–291, 1995.

[102] Anders Green, Helge Huttenrauch, and K Severinson Eklundh. Applying the wizard-of-oz framework to cooperative service discovery and configuration. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 575–580. IEEE, 2004.

[103] Anders Green and Kerstin Severinson-Eklundh. Task-oriented dialogue for cero: a user-centered approach. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, pages 146–151. IEEE, 2001.

[104] Zenzi M Griffin. Why look? reasons for eye movements related to language production. *The interface of language, vision, and action: Eye movements and the visual world*, pages 213–247, 2004.

[105] Zenzi M Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological science*, 11(4):274–279, 2000.

[106] Emile Halphen. *Des Lésions traumatiques de l'oreille interne*. 1910.

[107] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500, 2010.

[108] J Hansen and Vaishnevi Varadarajan. Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(2):366–378, 2009.

[109] John HL Hansen and Sahar E Bou-Ghazale. Robust speech recognition training via duration and spectral-based stress token generation. *Speech and Audio Processing, IEEE Transactions on*, 3(5):415–421, 1995.

[110] Laurence R Harris, Laurence Roy Harris, and Michael Jenkin. *Vision and action*. Cambridge University Press, 1998.

[111] Jeffrey J Hendrickson. Performance, preference, and visual scan patterns on a menu-based system: implications for interface design. In *ACM SIGCHI Bulletin*, volume 20, pages 217–222. ACM, 1989.

[112] Panikos Heracleous, Pierre Badin, Gérard Bailly, and Norihiro Hagita. Exploiting multimodal data fusion in robust speech recognition. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 568–572. IEEE, 2010.

[113] Panikos Heracleous, Miki Sato, Carlos T Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Speech production in noisy environments and the effect on automatic speech recognition. In *International Congress of Phonetic Sciences, Hong Kong, China*, pages 855–858, 2011.

[114] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990.

[115] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (rasta-plp). In *Second European Conference on Speech Communication and Technology*, 1991.

[116] Kris Hermus, Patrick Wambacq, et al. A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Applied Signal Processing*, 2007(1):195–195, 2007.

[117] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[118] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[119] HG Hirsch and C Ehrlicher. Noise estimation techniques for robust speech recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 153–156. IEEE, 1995.

[120] David Holman. Gazetop: interaction techniques for gaze-aware tabletops. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 1657–1660. ACM, 2007.

[121] Davis Howes. On the relation between the intelligibility and frequency of occurrence of english words. *The Journal of the Acoustical Society of America*, 29:296, 1957.

[122] C. W. Hsu, C. C. Chang, and C. J. Lin. *A practical guide to support vector classification*. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2003.

[123] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[124] Yi Hu and Philipos C Loizou. Subjective comparison of speech enhancement algorithms. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[125] Yi Hu and Philipos C Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *The Journal of the Acoustical Society of America*, 122:1777, 2007.

[126] Yi Hu and Philipos C Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7):588–601, 2007.

[127] Fu Jie Huang and Tsuhan Chen. Consideration of lombard effect for speechreading. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pages 613–618. IEEE, 2001.

[128] Hung-Hsuan Huang, Aleksandra Cerekovic, Igor S Pandzic, Yukiko Nakano, and Toyoaki Nishida. Toward a multi-culture adaptive virtual tour guide agent with a modular approach. *AI & society*, 24(3):225–235, 2009.

[129] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 257–264. ACM, 2003.

[130] JJ Humphries and PC Woodland. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In *Proc. Eurospeech*, volume 5, pages 2367–2370, 1997.

[131] Thomas E Hutchinson, K Preston White Jr, Worthy N Martin, Kelly C Reichert, and Lisa A Frey. Human-computer interaction using eye-gaze input. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(6):1527–1534, 1989.

[132] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1477–1480. ACM, 2004.

[133] Fumitacla Itakura and Shuzo Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan*, 53:36–43, 1970.

[134] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(1):67–72, 1975.

[135] Koji Iwano, Tomoaki Yoshinaga, Satoshi Tamura, and Sadaoki Furui. Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1):4–4, 2007.

[136] Rukmini Iyer, Mari Ostendorf, and Marie Meteer. Analyzing and predicting language model improvements. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 254–261. IEEE, 1997.

[137] R. J. K. Jacob and K. S. Karn. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. *The Mind's eye: Cognitive The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–603, 2003.

[138] Robert JK Jacob. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 11–18. ACM, 1990.

[139] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134, 2007.

[140] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and Joachim Quantz. *Dialogue acts in VERBMOBIL*. Citeseer, 1995.

[141] Frederick Jelinek, L Bahl, and R Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *Information Theory, IEEE Transactions on*, 21(3):250–256, 1975.

[142] Ljubomir Josifovski. *Robust automatic speech recognition with missing and unreliable data*. PhD thesis, Citeseer, 2002.

[143] Anna Judica, Maria De Luca, Donatella Spinelli, and Pierluigi Zoccolotti. Training of developmental surface dyslexia improves reading performance and shortens eye fixation duration in reading. *Neuropsychological Rehabilitation*, 12(3):177–197, 2002.

[144] J-C Junqua, Steven Fincke, and Ken Field. The lombard effect: A reflex to better communicate with others in noise. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 4, pages 2083–2086. IEEE, 1999.

[145] J-C Junqua and Hisashi Wakita. A comparative study of cepstral lifters and distance measures for all pole models of speech in noise. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 476–479. IEEE, 1989.

[146] Jean-Claude Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93:510, 1993.

[147] Jean-Claude Junqua. A duration study of speech vowels produced in noise. In *Third International Conference on Spoken Language Processing*, 1994.

[148] Jean-Claude Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, 20(1):13–22, 1996.

[149] Daniel Kahneman. Attention and effort. 1973.

[150] Ozlem Kalinli, Michael L Seltzer, and Alex Acero. Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3825–3828. IEEE, 2009.

[151] Sunil Kamath and Philipos Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE international conference on acoustics speech and signal processing*, volume 4, pages 4164–4164. Citeseer, 2002.

[152] Melih Kandemir, Veli-Matti Saarinen, and Samuel Kaski. Inferring object relevance from gaze in dynamic scenes. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 105–108. ACM, 2010.

[153] Z. A. Kapoula. The influence of peripheral preprocessing on oculomotor programming in a scanning task. *Eye movements and psychological functions: International views*, pages 101–114, 1983.

[154] Pawel Kasprowski and Józef Ober. Eye movements in biometrics. In *Biometric Authentication*, pages 248–258. Springer, 2004.

[155] Manpreet Kaur, Marilyn Tremaine, Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo, and Chandra Sekhar Mantravadi. Where is it? event synchronization in gaze-speech input systems. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 151–158. ACM, 2003.

[156] S Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation*, 15(7):1667–1689, 2003.

[157] A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22–63, 1967.

[158] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester. Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121:723, 2007.

[159] Brian ED Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1):117–132, 1998.

[160] Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 187–190. ACM, 2010.

[161] Katrin Kirchhoff, Gernot A Fink, and Gerhard Sagerer. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3):303–319, 2002.

[162] Dennis H Klatt. Review of the arpa speech understanding project. *The Journal of the Acoustical Society of America*, 62:1345, 1977.

[163] Jeff Klingner. Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 275–282. ACM, 2010.

[164] Jeff Klingner. Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 275–282, New York, NY, USA, 2010. ACM.

[165] Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing*, pages 41–50. Springer, 1990.

[166] Paul A Kolers, Robert L Duchnicky, and Dennis C Ferguson. Eye movement measurement of readability of crt displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 23(5):517–527, 1981.

[167] David B Koons, Carlton J Sparrell, and Kristinn Rr Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. *MIT Press: Menlo Park, CA*, pages 257–276, 1993.

[168] Xerxes P Kotval and Joseph H Goldberg. Eye movements and interface component grouping: an evaluation method. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 486–490. SAGE Publications, 1998.

[169] Ulrich H-G Kreßel. Pairwise classification and support vector machines. In *Advances in kernel methods*, pages 255–268. MIT Press, 1999.

[170] Roland Kuhn and Renato De Mori. A cache-based natural language model for speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(6):570–583, 1990.

[171] Solomon Kullback. *Information theory and statistics*. Courier Dover Publications, 1968.

[172] Michael Land, Neil Mennie, Jenny Rusted, et al. The roles of vision and eye movements in the control of activities of daily living. *PERCEPTION-LONDON-*, 28(11):1311–1328, 1999.

[173] Harlan Lane and Bernard Tranel. The lombard sign and the role of hearing in speech. *Journal of Speech, Language and Hearing Research*, 14(4):677, 1971.

[174] Kai-Fu Lee. On large-vocabulary speaker-independent continuous speech recognition. *Speech communication*, 7(4):375–379, 1988.

[175] CJ Leggetter and PC Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech and language*, 9(2):171, 1995.

[176] R Leonard. A database for speaker-independent digit recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 328–331. IEEE, 1984.

[177] Gregor Leusch, Nicola Ueffing, Hermann Ney, et al. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pages 33–40. Citeseer, 2003.

[178] Stephen E Levinson. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 1983.

[179] Junfeng Li, Shuichi Sakamoto, Satoshi Hongo, Masato Akagi, and Yôiti Suzuki. Two-stage binaural speech enhancement with wiener filter for high-quality speech communication. *Speech Communication*, 53(5):677–689, 2011.

[180] Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[181] Hsuan-Tien Lin and Chih-Jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. *submitted to Neural Computation*, pages 1–32, 2003.

[182] Richard P Lippmann. Review of neural networks for speech recognition. *Neural computation*, 1(1):1–38, 1989.

[183] Pierre Lison. A salience-driven approach to speech recognition for human-robot interaction. In *Interfaces: Explorations in Logic, Language and Computation*, pages 102–113. Springer, 2010.

[184] Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang. Feature selection with dynamic mutual information. *Pattern Recognition*, 42(7):1330–1339, 2009.

[185] Philipos C Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):47–56, 2011.

[186] Bruce Lowerre. The harpy speech understanding system. In *Readings in speech recognition*, pages 576–586. Morgan Kaufmann Publishers Inc., 1990.

[187] François Lux, Anna Mignot, Pierre Mowat, Cédric Louis, Sandrine Dufort, Claire Bernhard, Franck Denat, Frédéric Boschetti, Claire Brunet, Rodolphe Antoine, et al. Ultrasmall rigid particles as multimodal probes for medical applications. *Angewandte Chemie International Edition*, 50(51):12299–12303, 2011.

[188] Paul P Maglio, Rob Barrett, Christopher S Campbell, and Ted Selker. Suitor: An attentive information system. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 169–176. ACM, 2000.

[189] P. Majaranta and K.J. Räihä. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, page 22. ACM, 2002.

[190] Marc Marschark, Dominique LePoutre, and Linda Bement. *Mouth movement and signed communication*. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998.

[191] Sven C Martin, Jörg Liermann, and Hermann Ney. Adaptive topicdependent language modelling using word-based varigrams. In *Proc. Eurospeech97*, 1997.

[192] Thomas B Martin, AL Nelson, and HJ Zadell. Speech recognition by feature-abstraction techniques. Technical report, DTIC Document, 1964.

[193] David Maulsby, Saul Greenberg, and Richard Mander. Prototyping an intelligent agent through wizard of oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 277–284. ACM, 1993.

[194] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. The ami meeting corpus. In *Proceedings*

of the 5th International Conference on Methods and Techniques in Behavioral Research, volume 88, 2005.

[195] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[196] Eric Meisner, S Sabanovic, Volkan Isler, Linnda R Caporael, and Jeff Trinkle. Shadowplay: a generative model for nonverbal human-robot interaction. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 117–124. IEEE, 2009.

[197] Antje S Meyer, Astrid M Sleiderink, Willem JM Levelt, et al. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66(2):B25–B33, 1998.

[198] A.S. Meyer and C. Dobel. Application of eye tracking in speech production research. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, 2003.

[199] Matthew Middendorf, Grant McMillan, Gloria Calhoun, and Keith S Jones. Brain-computer interfaces based on the steady-state visual-evoked response. *Rehabilitation Engineering, IEEE Transactions on*, 8(2):211–214, 2000.

[200] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940. ACM, 2006.

[201] Harvey B Mitchell. *Multi-sensor data fusion: an introduction*. Springer, 2007.

[202] Vikramjit Mitra, Hosung Nam, Carol Y Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Articulatory information for noise robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):1913–1924, 2011.

[203] Shubham Mittal, Swati Vyas, and SRM Prasanna. Analysis of lombard and angry speech using gaussian mixture models and kl divergence. In *Communications (NCC), 2013 National Conference on*, pages 1–5. IEEE, 2013.

[204] Hansjörg Mixdorff, Ulrich Pech, Chris Davis, and Jeesun Kim. Map task dialogs in noise–a paradigm for examining lombard speech. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken*, pages 1329–1332, 2007.

[205] A-R Mohamed, Tara N Sainath, George Dahl, Bhuvana Ramabhadran, Geoffrey E Hinton, and Michael A Picheny. Deep belief networks using discriminative features for phone recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5060–5063. IEEE, 2011.

[206] Chafic Mokbel. *Reconnaissance de la parole dans le bruit: bruitage/débruitage*. PhD thesis, 1992.

[207] Chris Moore. Understanding the directedness of gaze: Three ways of doing it. *Infant and Child Development*, 15(2):191–193, 2006.

[208] Pedro J Moreno. *Speech recognition in noisy environments*. PhD thesis, Carnegie Mellon University, 1996.

[209] Pedro J Moreno, Bhiksha Raj, and Richard M Stern. A vector taylor series approach for environment-independent speech recognition. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 733–736. IEEE, 1996.

[210] Carlos H Morimoto and Marcio RM Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[211] KG Munhall, C Kroos, T Kuratate, J Lucero, M Pitermann, Eric Vatikiotis-Bateson, and H Yehia. Studies of audiovisual speech perception using production-based animation. In *Sixth International Conference on Spoken Language Processing*, 2000.

[212] Hy Murveit, Michael Cohen, Patti Price, Gay Baldwin, Mitch Weintraub, and Jared Bernstein. Sri's decipher system. In *Proceedings of the workshop on Speech and Natural Language*, pages 238–242. Association for Computational Linguistics, 1989.

[213] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68. ACM, 2009.

[214] Satoshi Nakamura, Hidetoshi Ito, and Kiyohiro Shikano. Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. 2000.

[215] Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa. Topic-dependent language model with voting on noun history. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(2):7, 2010.

[216] Dieter Nattkemper and Wolfgang Prinz. Saccade amplitude determines fixation duration: Evidence from continuous search. In *Eye movements: From physiology to cognition*. Elsevier, Amsterdam, 1987. editor: O'Regan, J. Kevin; editor: Levy-Schoen, A.

[217] Rajitha Navarathna, Patrick Lucey, David Dean, Clinton Fookes, and Sridha Sridharan. Lip detection for audio-visual speech recognition in-car environment. In *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, pages 598–601. IEEE, 2010.

[218] Jeannette G Neal and Stuart C Shapiro. Intelligent multi-media interface technology. *ACM SIGCHI Bulletin*, 20(1):75–76, 1988.

[219] Chalapathi Neti, Giridharan Iyengar, Gerasimos Potamianos, A Senior, and B Maison. Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)*, volume 3, pages 11–14, 2000.

[220] Gregg Norris and Eric Wilson. The eye mouse, an eye communication device. In *Bioengineering Conference, 1997., Proceedings of the IEEE 1997 23rd Northeast*, pages 66–67. IEEE, 1997.

[221] Sigrid Norris. *Analyzing multimodal interaction: A methodological framework*. Routledge, 2004.

[222] Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell. Evaluating look-to-talk: a gaze-aware interface in a collaborative environment. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 650–651. ACM, 2002.

[223] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference*, pages 101–1, 2011.

[224] Harry F Olson and Herbert Belar. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28:1072, 1956.

[225] John William Orillo, Roderick Yap, and Edwin Sybingco. Improved noise robust automatic speech recognition system with spectral subtraction and minimum statistics algorithm implemented in fpga. In *TENCON 2012-2012 IEEE Region 10 Conference*, pages 1–6. IEEE, 2012.

[226] Douglas OShaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[227] Mari Ostendorf, PJ Price, and Stefanie Shattuck-Hufnagel. The boston university radio news corpus. *Linguistic Data Consortium*, 1995.

[228] Sharon Oviatt. Multimodal interactive maps: designing for human performance. *Hum.-Comput. Interact.*, 12(1):93–129, March 1997.

[229] Sharon Oviatt. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 576–583. ACM, 1999.

[230] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[231] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. Linguistic adaptations during spoken and multimodal error resolution. *Language and speech*, 41(3-4):419–442, 1998.

[232] Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, et al. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human-computer interaction*, 15(4):263–322, 2000.

[233] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Referring Phenomena in a Multimedia Context and their Computational Treatment*, pages 1–13. Association for Computational Linguistics, 1997.

[234] Sharon L Oviatt. Multimodal signal processing in naturalistic noisy environments. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)*, volume 2, pages 696–699. Citeseer, 2000.

[235] Sharon L Oviatt and Karen Kuhn. Referential features and linguistic indirection in multimodal language. In *Proceedings of the International Conference on Spoken Language Processing*, volume 6, pages 2339–2342. Citeseer, 1998.

[236] SL Oviatt and Eric Olsen. Integration themes in multimodal human-computer interaction. In *Proceeding of the International Conference on Spoken Language Processing*, volume 2, pages 551–554. Citeseer, 1994.

[237] Kuldip Paliwal, Belinda Schwerin, and Kamil Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication*, 54(2):282–305, 2012.

[238] David S Pallett, Jonathan G Fiscus, William M Fisher, John S Garofolo, Bruce A Lund, and Mark A Przybocki. 1993 benchmark tests for the arpa spoken language program. In *Proceedings of the workshop on Human Language Technology*, pages 49–74. Association for Computational Linguistics, 1994.

[239] Hun Myoung Park. Comparing group means: t-tests and one-way anova using stata, sas, r, and spss. Technical report, Technical Working Paper. The University Information Technology Services (UITS) Centre for Statistical and Mathematical Computing, Indiana University, 2003.

[240] Christophe Patrick, Christophe Patrick Jan Van Bael, et al. Using the keyword lexicon for speech recognition. In *Department of Theoretical and Applied Linguistics, University of Edinburgh*. Citeseer, 2002.

[241] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[242] John Paulin Hansen, Allan W Andersen, and Peter Roed. Eye-gaze control of multimedia systems. *Advances in Human Factors/Ergonomics*, 20:37–42, 1995.

[243] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[244] Eric Petajan. Automatic lipreading to enhance speech recognition. 1984.

[245] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[246] James M Pickett. Effects of vocal force on the intelligibility of speech sounds. *The journal of the acoustical society of america*, 28:902, 1956.

[247] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005.

[248] Alexander Pollatsek, Keith Rayner, and David Balota. Inferences about eye movement control from the perceptual span in reading. *Attention, Perception, Psychophysics*, 40:123–130, 1986. 10.3758/BF03208192.

[249] Gerasimos Potamianos. Audio-visual automatic speech recognition and related bimodal speech technologies: A review of the state-of-the-art and open problems. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 22–22. IEEE, 2009.

[250] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[251] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23, 2004.

[252] DMW Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[253] Z. Prasov and J.Y. Chai. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481. Association for Computational Linguistics, 2010.

[254] Foster Provost and Pedro Domingos. Well-trained pets: Improving probability estimation trees. 2000.

[255] Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the fifteenth international conference on machine learning*, volume 445, 1998.

[256] Shaolin Qu and Joyce Y Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 193–200. ACM, 2006.

[257] Shaolin Qu and Joyce Y Chai. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics*, pages 284–291, 2007.

[258] Pernilla Qvarfordt and Shumin Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–230. ACM, 2005.

[259] L Rabiner, S Levinson, A Rosenberg, and J Wilpon. Speaker-independent recognition of isolated words using clustering techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(4):336–349, 1979.

[260] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[261] Bhiksha Raj and Richard M Stern. Missing-feature approaches in speech recognition. *Signal Processing Magazine, IEEE*, 22(5):101–116, 2005.

[262] P Rajasekaran, G Doddington, and J Picone. Recognition of speech under stress and in noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 733–736. IEEE, 1986.

[263] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422, 1998.

[264] K. Rayner. Eye movements in reading: Models and data. *Journal of eye movement research*, 2(5):1, 2009.

[265] K. Rayner, T.J. Smith, G.L. Malcolm, and J.M. Henderson. Eye movements and visual encoding during scene perception. *Psychological Science*, 20(1):6, 2009.

[266] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[267] Keith Rayner and George W. McConkie. What guides a reader's eye movements? *Vision Research*, 16(8):829 – 837, 1976.

[268] SR Research. http://www.sr-research.com/el_ii.html.

[269] D.C. Richardson and R. Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6):1045–1060, 2005.

[270] D.C. Richardson, R. Dale, and N.Z. Kirkham. The art of conversation is coordination. *Psychological Science*, 18(5):407, 2007.

[271] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition. In *icassp*, pages 81–84. IEEE, 1995.

[272] James Rossiter. *Multimodal intent recognition for natural human-robotic interaction.* PhD thesis, University of Birmingham, 2011.

[273] Deb Roy. Integration of speech and vision using mutual information. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 2369–2372. IEEE, 2000.

[274] Deb Roy and Niloy Mukherjee. Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, 19(2):227–248, 2005.

[275] Deb Roy, Bernt Schiele, and Alex Pentland. Learning audio-visual associations using mutual information. In *Integration of Speech and Image Understanding, 1999. Proceedings*, pages 147–163. IEEE, 1999.

[276] J Edward Russo and France Leclerc. An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, pages 274–290, 1994.

[277] Toshiyuki Sakai and Shuji Doshita. The phonetic typewriter. In *IFIP Congress*, volume 445, page 449, 1962.

[278] RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.

[279] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[280] Dario D Salvucci. Inferring intent in eye-based interfaces: tracing eye movements with process models. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 254–261. ACM, 1999.

[281] Dario D Salvucci and John R Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16(1):39–86, 2001.

[282] Ramesh R Sarukkai and Craig Hunter. Integration of eye fixation information with speech recognition systems. In *EUROSPEECH*, 1997.

[283] Christopher Schmandt and Eric A Hulteen. The intelligent voice-interactive interface. In *Proceedings of the 1982 conference on Human factors in computing systems*, pages 363–366. ACM, 1982.

[284] Jakub Segen and Senthil Kumar. Look ma, no mouse! *Communications of the ACM*, 43(7):102–109, 2000.

[285] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[286] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[287] R Shaw, E Crisman, A Loomis, and Z Laszewski. The eye wink control interface: using the computer to provide the severely disabled with increased flexibility and comfort. In *Computer-Based Medical Systems, 1990., Proceedings of Third Annual IEEE Symposium on*, pages 105–111. IEEE, 1990.

[288] Thomas B Sheridan. Function allocation: algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies*, 52(2):203–216, 2000.

[289] Elizabeth Shriberg, Raj Dhillon, Sonali Veda Bhagat, Jeremy Ang, and Hannah Carvey. *The ICSI meeting recorder dialog act (MRDA) corpus.* Defense Technical Information Center, 2004.

[290] Elizabeth Ellen Shriberg. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, 1994.

[291] Jaana Simola, Jarkko Salojärvi, and Ilpo Kojo. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9(4):237–251, 2008.

[292] Vasant Srinivasan and Robin R Murphy. A survey of social gaze. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 253–254. IEEE, 2011.

[293] M. Staudte and M.W. Crocker. Visual attention in spoken human-robot interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 77–84. ACM, 2009.

[294] Ralf Steuer, Jürgen Kurths, Carstens O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, 2002.

[295] Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming. Robust audio-visual speech recognition under noisy audio-video conditions. 2013.

[296] Petra-Maria Strauß, Holger Hoffmann, Wolfgang Minker, Heiko Neumann, Günther Palm, Stefan Scherer, Friedhelm Schwenker, Harald Traue, Welf Walter, and Ulrich Weidenbacher. Wizard-of-oz data collection for perception and interaction in multi-user environments. In *International Conference on Language Resources and Evaluation (LREC)*, 2006.

[297] Lu-ying Sui, Xiong-wei Zhang, Jian-jun Huang, and Bin Zhou. An improved spectral subtraction speech enhancement algorithm under non-stationary noise. In *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, pages 1–5. IEEE, 2011.

[298] Quentin Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):71–78, 1992.

[299] Kazuya Takeda, Atsunori Ogawa, and Fumitada Itakura. Estimating entropy of language from optimal word insertion penalty. In *Proceedings of Int. Conf. Spoken Language Processing*. Citeseer, 1998.

[300] Fotios Talantzis, Aristodemos Pnevmatikakis, and Lazaros C Polymenakos. Real time audio-visual person tracking. In *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, pages 243–247. IEEE, 2006.

[301] Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995.

[302] Rebecca M Todd, Deborah Talmi, Taylor W Schmitz, Josh Susskind, and Adam K Anderson. Psychophysical and neural evidence for emotion-enhanced perceptual vividness. *The Journal of Neuroscience*, 32(33):11201–11212, 2012.

[303] Pao Tsang-long, Chen Yu-te, and Yeh Jun-heng. Emotion recognition and evaluation from mandarin speech signals. *International Journal of Innovative Computing, Information and Control*, 4(7):1695–1709, 2008.

[304] Matthew Turk and Mathias Kölsch. Perceptual interfaces. *Emerging Topics in Computer Vision, Prentice Hall*, 2004.

[305] Christophe Van Bael and Simon King. An accent-independent lexicon for automatic speech recognition. International Congress of Phonetic Sciences, 2003.

[306] Maarten Van Segbroeck and Hugo Van Hamme. Advances in missing feature techniques for robust large-vocabulary continuous speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):123–137, 2011.

[307] Walter Van Summers, David B Pisoni, Robert H Bernacki, Robert I Pedlow, and Michael A Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917, 1988.

[308] A. Varga and H.J.M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[309] AP Varga and RK Moore. Hidden markov model decomposition of speech and noise. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 845–848. IEEE, 1990.

[310] Saeed V Vaseghi and Ben P Milner. Noise-adaptive hidden markov models based on wiener filters. In *Third European Conference on Speech Communication and Technology*, 1993.

[311] D. Vernon, G. Metta, and G. Sandini. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *Evolutionary Computation, IEEE Transactions on*, 11(2):151–180, 2007.

[312] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, pages 35–70, 2004.

[313] Roel Vertegaal, Connor Dickie, Changuk Sohn, and Myron Flickner. Designing attentive cell phone using wearable eyecontact sensors. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 646–647. ACM, 2002.

[314] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[315] TK Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1):52–57, 1968.

[316] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

[317] Paolo Viviani. Eye movements in visual search: cognitive, perceptual and motor control aspects. *Reviews of oculomotor research*, 4:353, 1990.

[318] Salah Werda, Walid Mahdi, and Abdelmajid Ben Hamadou. Lip localization and viseme classification for visual speech recognition. *arXiv preprint arXiv:1301.4558*, 2013.

[319] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[320] Jay G Wilpon, Lawrence R Rabiner, C-H Lee, and ER Goldman. Automatic recognition of keywords in unconstrained speech using hidden markov models. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(11):1870–1878, 1990.

[321] Ming Yan, Reinhold Kliegl, Hua Shu, Jinger Pan, and Xiaolin Zhou. Parafoveal load of word n+ 1 modulates preprocessing effectiveness of word n+ 2 in chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6):1669, 2010.

[322] A.L. Yarbus. *Eye movements and vision.* Plenum press, 1967.

[323] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1):23–43, 1998.

[324] Laurence R Young and David Sheena. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7(5):397–429, 1975.

[325] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*, volume 2. Citeseer, 1997.

[326] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[327] Qiaohui Zhang, K Go, A Imamiya, and Xiaoyang Mao. Designing a robust speech and gaze multimodal system for diverse users. In *Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference on*, pages 354–361. IEEE, 2003.

[328] Qiaohui Zhang, Atsumi Imamiya, X Mao, and K Go. A gaze and speech multimodal interface. In *Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference on*, pages 208–213. IEEE, 2004.

[329] Xu Zhang, Xiang Chen, Wen-hui Wang, Ji-hai Yang, Vuokko Lantz, and Kong-qiao Wang. Hand gesture recognition and virtual game control based on 3d accelerometer and emg sensors. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 401–406. ACM, 2009.

[330] Zhiwei Zhu and Qiang Ji. Eye and gaze tracking for interactive graphic display. *Machine Vision and Applications*, 15(3):139–148, 2004.

[331] Zhiwei Zhu and Qiang Ji. Novel eye gaze tracking techniques under natural head movement. *Biomedical Engineering, IEEE Transactions on*, 54(12):2246–2260, 2007.